Università
Ca'Foscari
Venezia

Master's degree in Economics and Finance

Final Thesis

A gravity model for Italian domestic tourism flows.

Supervisor
Professor Roberto Casarin

Assistant supervisor
Professor Dario Palumbo

Graduand
Carlos Rodríguez Ameal
883259

Academic Year
2020/2021

Acknowledgements

First, I would like to thank my supervisor Professor Roberto Casarin, and the assistant supervisor Professor Dario Palumbo, who has kindly corrected many of my mistakes and has shared with me very useful indications. I am also very grateful to Professor Marco Di Cataldo for having put me in contact with Dario and guiding me in my first steps of the project, when I had most doubts about it.

I must also thank my flatmates, Jota Maraca, Gonría and my parents, who have been there when I needed it and have acknowledged my concerns from a place of support and understanding. Finally, although I cannot mention them all, I am forever grateful to all my friends and family, who believe in me more than I often do.

Abstract

After decades of research applying the gravity model to study international tourism flows, some works are now are being conducted at a national scope. This work aims to contribute to this more recent trend by analysing the tourism flows among Italian regions for over a decade. The specification includes some classic determinants such as distance, population, income and prices; and others such as cultural institutions, crime rates and a measure of economic distance accounting for Linder's Hypothesis. The model uses a Pseudo-Poisson Maximum Likelihood estimation with time effects, due to being consistent with zero-valued flows, and following the recommendations in the current literature on Gravity Models. Some guidance on the treatment required by models based on pseudo-distributions is presented. The results are compared with those obtained from a classic Ordinary Least Squares after zero-valued flows are transformed adding 1 to them. They are in the greatest part in accordance to predictions and close to the results obtained by other authors, although the variables based on price and some others such as economic distance yield contradictory or ambiguous results. In order to obtain some clarification and as a robustness test, the Bayesian Model Averaging method is introduced and then applied to the data, reassuring the role of the classic gravity determinants and confirming Linder's Hypothesis.

Contents

List of Figures

Chapter 1

Introduction

Tourism is a key industry in Italy. In 2018 it was directly responsible for 5% of the national GDP in and 6% of jobs, while estimations of their indirect effect raise the figures to 13% of GDP and 15% of the jobs. These figures are only comparable with Spain out of all the big European economies, and figures of the previous decade showed a growth of tourism expenditure of around 9% a year. This growth has been mainly due to the increasing interest of non-European countries in Italy compared to other European destinations, performing especially well with Chinese tourists, thanks to the cultural and recreational offer of the country. Indeed, with 55 world heritage sites as of 2021, leading the global ranking of countries, and more hosting places than any other European country, Italian tourism had everything to boom in an otherwise modest economic panorama (see Petrella et al. (2019)).

The Covid-19 pandemic was a devastating hit to the tourism industry. With one of the earliest lock-downs in the globe, many businesses previously sustained by tourism seemed to be destined to close. Even when lock-downs were relieved, restrictions to international travels made it impossible for the tourism industry to recover. It is in this context that domestic tourism became the center of attention of public authorities as the best remedy to the otherwise gloom prospects and a lot of resources, economic but also mediatic, were destined to its promotion. The zenith of this new focus on national tourism was the endowment of 500€ to the Italian families with earnings under 40.000€ to spend on hotels and accommodations of the country. Before the pandemic, it would have been unthinkable that an industry with such a previously diversified clientele would depend on the nationals for its survival.

Thus, it could be safely stated that the necessity to model and understand domestic tourism is more imperious than ever. In this work I will analyse Italian domestic tourism flows, trying to identify which are the main determinants that influence them. In order to do so, I will follow the most common approach in current literature and use a panel-data gravity model specification. I will study the flows among Italian regions from 2008 until 2019, finishing right before the pandemic started. In order to account for zeros travel flows and problems arising from heterokedasticity I will use a Pseudo-Maximum Likelihood Poisson (PPML) estimator with time effects, although I also present the results of an Ordinary Least Squares (OLS) estimation for contrast. To obtain a comprehensive view on the

tourism determinants I will include a wide set of explanatory variables, ranging from classic determinants as distance, population, GDP per capita and prices, to crime rate, cultural institutions or economic distance à la Linder. To respond to possible concerns derived from the inclusion of a relatively high number of variables I have run a Bayesian Model Averaging as a robustness test.

Although domestic tourism has received much less attention than international tourism in the literature, there already exists a paper, Massidda and Etzo (2012), where Italian domestic is studied and I will be often reference it and use it to compare my results. Still, my approach is innovative in a series of points. First, it is one of the minority of works which mainly base on a economic theory derivations of the gravity model for tourism to select the explanatory variables. Furthermore, I also study Linder's hypothesis in the case of domestic tourism, being able to compare both approaches and draw conclusions on their validity. Second, I show how to empirically work with a PPML specification. As all models based on pseudo-distributions, the PPML does not really assume an underlying distribution for the data, so the goodness-of-fit tests and the diagnostics of the model need to be interpreted with caution and not all usual procedures can be validly performed. Although there are strong cases from papers supporting the use of this pseudo-model in the particular case of the gravity model, there is still a certain degree of confusion in the literature regarding this and other pseudo-models and their differences with the regular generalised linear models. I hope that my work will offer some guidance on the use of (at least Poisson) pseudo distribution estimations for the gravity model. Finally, I propose a Bayesian Model Averaging as a robustness test that is compatible with the PPML. This relatively new technique allows to an alternative interpretation of significance of the variables that can be more intuitive for the reader, as well as being inherently less limited than the classic frequentist approach to model estimation. Its use is particularly useful for research where the effect of several determinants is to be studied and researchers want to take an agnostic approach to the question of the correct specification, as it is the case of this work.

The results are generally in accordance to previous conclusions of papers working on tourism and domestic tourism in particular. The classic variables: distance, population and GDP per capita, show to be strong determinants of domestic tourism flows in Italy. In particular, distance seems to be a stronger determinant than in the case of international tourism, showing that the gravity model is more useful to model national than international flows, as some other researchers have claimed. Despite the success of the gravity variables, Linder's effect seems to be also an important factor, suggesting that both approaches can actually be complementary, challenging the common perception of their incompatibility. Besides, some room of action is left for public authorities: results point towards a robust positive effect of cultural attractions, whose development can be a tool to increase the attractiveness of a regions, while reducing criminality should also have a positive effect in the incoming tourism. Prices in the other case do not seem to deter tourism according to the model, but actually high prices may be positively correlated to tourism appeal.

Regarding the structure of the work, it is formed by five chapters. The remaining part of this chapter will be dedicated to reviewing the most important literature on which I

have based. I introduce the reader to the founding works of the gravity model, its current state of the art and its application in the field of tourism. Chapter 2 will correspond to the presentation of my model, I explain the theory behind the PPML but I also present the OLS specification that I include for comparison, and I motivate the determinants that I use. In Chapter 3 I present the data and the results of the model, along with a discussion of them. In Chapter 4 I quickly introduce the reader to the most important concepts of Bayesian statistics, I present the theoretical framework of the BMA and show the results of its implementation. Finally, Chapter 5 will collect the conclusions.

1.1 Literature review

Gravity models were first conceived in the first half of the 20th century during a wave of economic modeling based on contributions by other disciplines, namely physics. Some other remarkable examples include the application of Zipf's law to demographics or Reilly's law of retail gravitation. However, the main work of interest for this work was developed by astrophysicist John Quincy Stewart (see Stewart (1948)). Stewart used Newton's formula of universal gravitation to model migration flows among cities by substituting physical mass with population[1] . Thus, the "demographic force" would be given by:

$$F = \frac{GN_1N_2}{d^2}, \tag{1.1}$$

where N_1 and N_2 are the populations of the localities of study, d is the distance between them and G is the equivalent of the gravitational constant whose value has to be estimated. Stewart did not stop there, going beyond with the physical analogies and also translated to the demographic context other related concepts such us energy or potential.

Further contributions were made by the American economist Walter Isard. Basing on previous works, including Stewart (1948), in Isard (1954) he proposed that a gravity model be applied to international flows of commodities, which started the popular practice of applying the gravity model to international trade. The formulation that is used today is essentially similar to the one proposed by Dutch economist Jan Tinbergen. In Tinbergen (1962) he proposed a model where trade from country i to country j, X_{ij}, was proportional to the economic masses of the countries, M_i and M_j, and inversely proportional to the distance between them, d_{ij}. Allowing for more generality, distance did not necessarily enter the equation squared, but with an unknown exponent that had to be determined, and it is joined by a dummy variable, N_{ij}, a dummy showing if the pair of countries shared a border. Finally, he added other variables accounting for trade preference, in particular two other dummy variables, P_C and P_B, indicating if the countries belonged to a supranational union, namely the Commonwealth of Nations or Benelux respectively. The specification is then the

[1]Actually Stewart referred to "molecular mass". This was a modified population variable, since simple population could not explain the big differences among countries of equal size.

following:

$$X_{ij} = \frac{G M_i^{a_2} M_j^{a_3} N_{ij} P_{Cij} P_{Bij}}{d^{-a_1}} \lambda_{ij}, \tag{1.2}$$

where λ_{ij} is an error term with a log-normal distribution. Applying logarithms to both inequalities leads to the equivalent linear formulation:

$$\ln X_{ij} = \ln G + a_1 \ln d + a_2 \ln M_i + a_3 \ln M_j + N_{ij} + P_{Cij} + P_{Bij} + \epsilon_{ij}, \tag{1.3}$$

where $\epsilon_{ij} = \ln \lambda_{ij}$ is now assumed to be normally distributed. Applying logarithms to the variables the coefficients gained a new interpretation as the elasticities of trade with respect to each explanatory variable. The elasticities, which were assumed to be constant, were estimated by means of a regular OLS cross-country regression, and only non-zero trade flows were included in the analysis since the log-log specification cannot account for zero-flows. Note that allowing for different elasticities depending on the role of the country (origin or destination) allows for different estimated values depending on the direction of the flow, which is something desired as empirically inflows and outflows are usually fairly different in magnitude.

Tinbergen obtained coefficient estimates for GNP and distance with the expected sign and that were statistically significant. Although there were several specification issues with this simple model, it was promising enough to foster a great amount of interest in what was named the Trade Gravity model, which became the most common application of the gravity models and whose popularity continues until today. Subsequent research focused on trying to better explain differences in preferences among countries. The parameter of gravitational attraction, G, actually variables from pair to pair of countries and since preferences are not exogenous, models that do not account for it yield biased estimations. For example, estimating the effect of entering in a supranational organisation such as the Commonwealth (taking Tinbergen's specification) can be challenging as this variable can correlate with others that he did not include, but that may also have an effect such as speaking the same language or sharing the same principal religion. The application of different methods to sort this bias out range from trying to include as many explanatory variables as possible, to giving preference to panel-data analysis over cross-sectional ones.

While one of the main appeals of the gravity model is that, contrary to competing models, its application does not require modelling demand and supply, nor dynamics are usually involved, the initial lack of a theoretical foundation made some authors wary of its application. Indeed, after the boom of the gravity model, most authors focused on its use for empirical research, shutting eyes to the lack of theoretical justification. At the same time, some other competing and better theoretically founded models appeared. Perhaps the most influential one is the one developed by Swedish economist and minister of trade H. M. S. Linder, who modeled trade flows basing on a demand-supplied approach in his paper Linder (1961). The main consequence of Linder's work is that trade between two countries should increase as their income per capita gets closer. This would happen as factors' prices are more similar among the groups of developed and developing countries than between them, fostering the production of more similar items in each group and therefore specialising in

products of a shared appreciation among them. In particular, rich countries have a more capital-intensive industry, producing products of higher quality that are consumed mainly by them or by other developed countries. This theorised phenomenon is currently known as *Linder's Hypothesis*, and it has been tested in several occasions in the case of trade and also in the particular case of tourism. Results are not determinant, for example, Keum (2010) found that including a variable measuring the absolute distance between GDP per capita to test its influence on trade flows with Korea yielded a negative coefficient estimate if the dependent variable was exports (as it would be expected), but it became positive when the dependent variable was imports.

It would be necessary to wait until in the 1980's some authors that noticeable contributions developing micro-economic models that could sustain the gravity equation. One of the first and most notorious attempts in this regard comes from Anderson (1979), who was able to derive the multiplicative formulation of the gravity model from a theoretical framework, under some economic assumptions such as product differentiation by place of origin and Constant Elasticity of Substitution expenditures. Since then, several decades of theoretical work have led to an abundance of models that provide the much needed backing that Tingenberg's formulation wanted. For a discussion on the main different approaches see Anderson (2011). Note that these derivations arrive to the multiplicative equation of the gravity model, which since Tinbergen (1962) takes a stochastic approach when applied to trade, being too complex a phenomenon to attempt to explain it with a deterministic equation. Therefore, the gravity model formulation should be understood as the conditional average of flows given the certain characteristics depending on the country of origin, destination or on both at the same time.

The focus today has moved towards finding the right specification as to avoid all the possible sources of bias from which the gravity model can be affected. One of the main concerns that have been pointed out and that has led to much concern on past gravity model specifications refers to the Multilateral Resistance bias. This bias arises when to explain flow from country i to country j, specifications only count on variables depending on i and j alone (e.g. GDP of each country), and on both i and j at the same time (e.g. distance between countries), but not on the characteristics of the other countries. In other words, all variables are either unilateral or bilateral in nature. However, in a very influential paper, Anderson and Van Wincoop (2003), and building on the previously mentioned paper by Anderson, the authors pointed out the necessity of including a multilateral resistance term. This term would reflect the influence on country i of the obstacles for trade with other countries on trade with country j. If a different country suddenly becomes more attractive for country i, trade with country j may be negatively affected, even if all bilateral and unilateral variables remained the same. The different methods that authors have followed to reduce this bias depend on the variables used. First, some do not address this issue at all, or if they do, they admit their lack of empirical strategy to overcome it. Another group of authors use origin and destination region fixed effects to account for it, although very rarely they are time-varying as it was indicated in Anderson and Van Wincoop (2003). Finally, the last group includes a separate variable that tries to represent the relative resistance of trade from seller i to buyer j compared to every other possible destination, usually using price,

distance, GDP or a combination of them.

Also to avoid miss-specification, some more other estimation techniques have been competing with the OLS specification, mainly Generalised Linear Models. In particular, models based on the Poisson or in the Negative Binomial distribution have been increasingly popular as previous works based in normal errors have been more contested. One strong case against the OLS specification came from Silva and Tenreyro (2006), who affirmed that the OLS specification suffered from bias under heterokedasticity of the errors, and proposed a Poisson-Maximum Likelihood estimator instead. In a very recent paper, Tyazhelnikov and Zhou (2020) added that, leaving this bias aside, in the case that elasticities are not constant across regions, the interpretation of an OLS specification is different from that of a Poisson. In any case, some other interesting properties of the mentioned generalised estimators, namely the ability to handle zero-valued flows, have made them the preferred specification in a multitude of works.

Parallel to these debates, the use of the gravity model grew more and more prevalent in a diversity of fields. In this work it will be applied to model tourism, which can considered as a specific type of trade of a service among regions. Today, gravity models for tourism benefit from a number of empirical works corroborating its adequacy and a few theoretical derivations of their own that can serve as a basis for researchers. In this work this role will be played by Morley et al. (2014), who derived a gravity model for tourism from demand theory in a particularly concise way, but also note Cochrane (1975) contribution to the more general case of trips. Furthermore, after applying the gravity model to a data set that differentiates services trade from goods trade, Kimura and Lee (2006) showed that trade of services is better predicted by the gravity model than trade of goods, in the sense that it shows stronger and more significant effects of GDP and distance on trade. This explains the ever stronger preeminence of the gravity in tourism studies, which has also permeated institutional research of tourism determinants (e.g. Culiuc (2014) for the IMF). Yet, Mayo et al. (1988) contested the use of the gravity model to model international tourism, since the relation between tourism and distance might be less straightforward than in the case of trade, as exotic destinations can be more attractive than nearby ones.

In particular, this work is concerned with the study of domestic tourism flows, where exostism is likely to play a much smaller role than in the international case. Still, in the already existent paper studying domestic tourism flows in Italy, Massidda and Etzo (2012), it is found that distance plays a lesser role than usual estimates for the international case. This result was not corroborated by Priego et al. (2015), where Spanish domestic tourism flows were studied, who obtained an elasticity for distance around ten times more negative than Massidda and Etzo, 2012. Whether distance plays a bigger or a smaller role when flows are among closer regions is a question to which there is still no clear answer and to which this project will try to contribute.

With the maturity of the gravity model and the diversification of its applications, many contemporary approaches mix the classic specification with new techniques. Indeed, the proven strength of the gravity formulation along with its simplicity makes it especially attractive for its implementation with the most modern statistical trends. One example of

a especial pertinence for this work is the rise of the application of Bayesian tools to the models. Most works incorporate them in the model specification such as Ranjan and Tobias (2007), who centred on the role of institutions on trade by specifying a tobit model with a bayesian approach, or Congdon (2000), who based on bayesian statistics to allow for structural variation in the distance coefficient. However, other works have applied it to reflect model uncertainty, following a Bayesian Model Averaging procedure (see Beck et al. (2017a) and Chen et al. (2018)).

Thus, the popularity of the gravity model is explained by its good performance after decades of applications to model first trade, and then any possible flow where distance could play a factor. But also due to its simplicity and the versatility of its formulation, which allows for influences from the most modern techniques. For a deeper review of the history of the gravity model and the current discussions on its use see De Benedictis and Taglioni (2011) and Shahriar et al. (2019).

Chapter 2

The model

2.1 The Ordinary Least Squares model

Current analysis on tourism flows applying the gravity model does not discriminate between leisure and professional travel, encompassing everything in the same category: a service that is consumed by individuals since they benefit from it. As such, the specifications that the gravity model take when applied to tourism are the same as the ones when it is applied to trade, and a general linear formulation can be expressed in the same way that Tinbergen did (equation 1.3), but substituting trade volume by arrivals.[1]. For a panel-data model, the specification would be as follows:

$$E(X_{ijt}|Z) = G \prod_k ZO_{kit}^{\zeta_k} \prod_l ZD_{ljt}^{\eta_l} \prod_m ZOD_{mijt}^{\theta_m}, \tag{2.1}$$

where Z is the set of all explanatory variables and where every characteristic k of country i at time t as a supplier is stacked in vector ZO_{it} (e.g. GDP of the exporter), every characteristic l of country j at time t as a consumer is stacked in ZO_{jt} (e.g. GDP of the importer) and every bilateral variable m accounting for the accessibility of country i to j is stacked in ZOD_{ijt} (e.g. distance between countries). Adding a log-normal error term, λ_{ijt}, as a factor as in equation 1.2 and applying logarithms to both sides results in the linear OLS specification:

$$\ln X_{ijt} = \alpha + \sum_k \zeta_k \ln ZO_{kit} + \sum_l \eta_l \ln ZD_{ljt} + \sum_m \theta_m \ln ZOD_{mijt} + \epsilon_{ijt}, \tag{2.2}$$

In the linear expression the error term, $\epsilon_{ijt} = \ln \lambda_{ijt}$, is normally distributed and $\alpha = \ln G$ is the constant.

[1]The flow value can be defined as the actual flow of travellers from one region to another when the time spent abroad does not matter, in which case it is referred to as arrivals. If instead the flow is equal to the number of days spent by people from region i in region j it is referred to as stays. The first variable is the most commonly used since theoretical derivations do not usually consider different travel durations and their determinants are less clearly established

Note that in this equation the origin country of the trade flow, i, was the supplier of the trade, and the destination country, j, the consumer. In the case of tourism we call flow the number of people visiting country j from country i, implying that now the destination is the supplier (of the tourism service) and the origin the consumer. Since referring to a flow as X_{ij} with i and j noting "origin" and "destination" instead of "supplier" and "consumer" is more intuitive it will be kept like this, and the general structural symmetry of the gravity equation still allows for an equivalent formulation to that of equation 1.3.

There are some problems with this specification. First, in the case that there are flows with a value equal to zero, it is clear that the log-log specification cannot hold. However, the zero flows can neither be eliminated as its appearance is not random, but reflect too strong barriers for trade to occur. A possibility is to simply add one to every flow, so that it the log-log specification can be applied. I have used this as one of the possible specifications. As a more consistent solution (the arbitrarity of adding one to each flow does not make this option very trustworthy) there have been different non-linear models that have been applied substituting the naif Ordinary Least Squares. The most common models include the Poisson Pseudo-Maximum Likelihood (PPML), the Negative Binomial Pseudo-Maximum Likelihood (BNPML) or the Tobit model. If the occurrence of zeros is too frequent, the first two can be modified to the Zero-Inflated Poisson Pseudo-Maximum Likelihood (ZIPPML) or the Zero-Inflated Negative Binomial Pseudo-Maximum Likelihood (ZINBPML). A lot has been written on the adequacy of each model and its performance, but there is still no consensus on the best choice, which also should attend the characteristics of the data involved.

A related issue associated with the assumption of a log-normal model and an OLS estimation originates when heterokedasticity[2] is present in the model, which is very often the case. Even if we assume that the error term in the multiplicative formulation is uncorrelated with the explanatory variables, this does not imply that it continues to be so after applying a logarithm in the linear specification. From Jensen's inequality we know that $E(\ln \lambda_{ijt}|Z) \neq \ln E(\lambda_{ijt}|Z)$, where λ_{ijt} is the error term in the multiplicative expression and Z represents the set of explanatory variables. And the difference is a function of the higher moments of $\lambda_{ijt}|Z$. Thus the presence of heterokedasticity of the log-normal terms, the error terms in the linear equation will be correlated with the explanatory variables and the specification will not yield consistent estimators. Silva and Tenreyro (2006) address this issue and propose the Poisson Pseudo-Maximum Likelihood specification as a solution. Although the PPML uses a Poisson likelihood function and the Poisson distribution assumes equidispersion (the conditional variance is equal to the conditional mean), the authors show that the estimation is still consistent in the case of over dispersion. It is even consistent for more general cases in which data is not count, and when zeros are more common than predicted by a Poisson model. Basing on this, I have also included a PPML specification to the model,

[2]A vector of random variables is said to present heterokedasticity when the variance is not constant across the elements of the vector. In a regression model, the error terms are said to be homokedastic when they present the same variance regardless of the value of the explanatory variables, if not they are said to be homokedastic. Using the notation of equation 2.1, a sufficient condition for homokedasticity is that $V(\lambda_{ijt}|Z_{ijt}) = \sigma^2$, with σ^2 constant.

which will be the preferred specification in case of discordance with the OLS.

A final issue that is shared with most panel-data econometric regression models is the inclusion of individual effects on the specification. Again, much has been written and argued on this regard. As it was explained before when discussing Tingenberg's paper, the omitted variable bias is a serious issue that concerns most studies based on the gravity model, and the inclusion of individual effects under a panel-data framework is one of the most effective ways to account for it. Individual effects specifications can assume independence of the effects with the error term, in which case they are referred as random effects as they behave as an independent random variable, or fixed effects, when that assumption cannot be made since otherwise it would yield biased estimates. The currently most common specification includes a set of fixed effects for the origins and another for the destinations, as in the following formulation:

$$\ln X_{ijt} = \alpha + \sum_k \zeta_k \ln ZO_{kit} + \sum_l \eta_l \ln ZD_{ljt} + \sum_m \theta_m \ln ZOD_{mijt} + \gamma_i + \delta_j + \epsilon_{ijt}. \quad (2.3)$$

This specification allows to consistently estimate the elasticities of interest with a simple OLS (although not in the case of heterokedasticity) and does not require assuming their uncorrelation with the omitted variables, which do not need to be proxied by including more variables than the one of interest making data requirements diminish. Furthermore, Anderson and Van Wincoop (2003) show that the inclusion of origin and destination unilateral fixed effects can account for the multilateral resistance term, although its changing nature requires the effects to be time varying. Despite their benefits, their inclusion can originate several incompatibilities with the estimation of the variables of study. If for example a variable is only origin-dependent and does not vary over time, perfect collinearity prevents the estimation of its elasticity. Even when variables are time-varying, if changes over time are not significant enough estimations can be seriously affected. Only when the study revolves over a variable dependent on both the origin and destination this specification can be safely performed. On the other hand, the inclusion of pair fixed effects is not common, as it precludes the study of the role of distance due to perfect collinearity and of any other variable of interest that is not time-varying. On top of that, if the data is too massive the estimation of whichever specification of fixed effects can be computationally impossible, and most common methods for estimation in this case such us the Within or the First-difference estimations do not allow for the study of any time-constant explanatory variable. Although in this case the Hausman and Taylor estimator can be used, as it is not consistent with the PPML estimation, I have opted for not including origin and destination fixed effects in neither specification, and take advantage of it to include variables time-constant unilateral variables instead.

2.2 The Poisson-Pseudo Maximum Likelihood model

As advanced before, I will follow a PPML specification in order to be allowed to include the flos equal to zero and to avoid the log-normal inconsistency under heterokedasticity. A

Poisson regression for count data is defined by the conditional density:

$$Pr(X_{ijt}|Z_{ijt}) = \frac{\exp\{-\mu_{ijt}\}\mu_{ijt}^{X_{ijt}}}{X_{ijt}!}, \tag{2.4}$$

with:

$$\mu_{ijt} = \exp\{Z_{ijt}\beta\}. \tag{2.5}$$

The equi-dispertion characteristic of the Poisson model refers to the fact that: $\mu_k = E(x_k|z_k) = V(x_k|z_k)$. The estimate of the vector of coefficients comes from the maximisation of the log-likelihood function:

$$\ln \mathcal{L}(\beta) = \sum_{i,j,t} -\mu_{ijt} + X_{ijt}(Z_{ijt}\beta) - \ln X_{ijt}! \tag{2.6}$$

In order to find these estimates we set the score vector (the gradient of the log-likelihood) equal to zero and obtain the first-order conditions:

$$s(\beta) = \sum_{i.j.t}[X_{ijt} - \exp\{Z_{ijt}\beta\}]Z_{ijt} = 0. \tag{2.7}$$

The Poisson estimator is the solution of this equation. The Hessian matrix of the log-likelihood is:

$$\frac{\partial \ln \mathcal{L}}{\partial\beta\partial\beta'} = \sum_{i,j,t} \exp\{Z_{ijt}\beta\}Z'_{ijt}Z_{ijt}, \tag{2.8}$$

The Poisson estimator is well behaved and it is well defined since the Hessian matrix of the log-likelihood is negative definite, therefore no more than one solution to 2.7 can exist. This also helps facilitate its numerical computation.

From the form of the first-order condition it is clear that β will be consistently estimated as long as $E(x_k|z_k) = \exp\{z_k\beta\}$. Thus, the data does not need to be Poisson distributed to use this method; in fact, it does not even need to be integer. If this is the case, the coefficients can still be estimated in the same manner as if it were Poisson-distributed, and the robust co-variance matrix of the residuals can be estimated regardless of the equi-dispersion restriction. This procedure is known as the Poisson Pseudo-Maximum Likelihood estimation, which in fact leads to the same coefficient estimates than the Poisson regression, but allows for heterokedasticity in a more general way. Instead of setting the conditional variance to be equal to the conditional mean, the covariance matrix of the residuals is directly computed using some robust estimator. More generally, a model based on a pseudo-distribution does not actually assume an underlying distribution of the data. It simply appropriates the first order condition derived from the density of a known distribution and uses it to model the data, hoping that the resulting estimates will be adequate enough. Thus, since they are not really based on a distribution, these models lack a density function that could permit the use of likelihood-based techniques, at least a priori. In particular, the PPML model does not have a likelihood function to compute goodness-of-fit measures such as Akaike's Information

Criterion[3] (AIC) or Bayes Information Criterion (BIC)[4]. These measures are commonly used in statistical research to make decisions over the set of variables that are included in a model. [5].

Still, models based on pseudo-distributions are very often analysed using the likelihood function of the origin distribution but in a slightly different way, in particular computing their deviance. In statistics, deviance is a fairly general concept that refers to a goodness-of-fit statistic applied to a model. In the context of a pseudo-maximum likelihood regression, there are two main "deviances" of interest: the null deviance and the residual deviance. Each of these statistics compare the likelihood of a different model with the one of the saturated model (where every observation has a separate parameter). The null deviance compares the likelihood of the model that only includes an intercept with the the saturated one, while the residual deviance studies the model with the specification of interest (often called the "proposed model"). Thus, the statistic in each case takes the form:

$$D(y;\hat{\mu}) = -2[\mathcal{L}(x) - \mathcal{L}(\hat{\mu})] \tag{2.9}$$
$$D(y;\bar{x}) = -2[\mathcal{L}(x) - \mathcal{L}(\bar{x})] \tag{2.10}$$

Here $\mathcal{L}(y)$ is the likelihood of the saturated model, $\mathcal{L}(\hat{\mu})$ that of the proposed model and $\mathcal{L}(\bar{x})$ that of the intercept-only model. To see a more formal definition of the deviance in the context of the Poisson regression see Liu (2019).

Since the saturated model will have the highest likelihood, the deviance is always positive, and the closer the deviance of a model is to zero the better the data is explained. On the other hand, the informative value of a model can be studied by comparing its deviance with the null one. In fact, under some conditions, the residual deviance derived from a Poisson model follows a χ^2 distribution provided that the model is well specified, so a goodness-of-fit test arises quite naturally (see Dunn and Smyth (2018) for a derivation and discussion on this test). However, if, like in this work, the model does not assume an underlying Poisson

[3]Akaike's information criteria is a measure of the quality of a model for a given data. Be k the number of the estimated parameters and $\mathcal{L}(\hat{\mu})$ the maximum of the likelihood function of the model, Akaike's information criterion (AIC) is defined as:

$$2k - 2\ln\mathcal{L}(\hat{\mu})$$

The lower the value the better the model fits the data, and overparametrisation is penalised by the inclusion of the term $2k$.

[4]An alternative to AIC, Bayes information criterion of a model with n observations, k variables and maximum likelihood of the mode $\mathcal{L}(\hat{\mu})$ for the likelihood function, is defined as:

$$BIC = k\ln n - 2\ln\mathcal{L}(\hat{\mu})$$

BIC is a more restrictive criterion than AIC, penalising more the inclusion of variables as the observations grow. Thus, the use of BIC leads to more parsimonious models than the use of AIC.

[5]A way of proceeding only basing on one of this criteria would consist in starting with a model including all potential variables of interest, and remove variables one by one trying to minimise the value of the chosen criterion. In an economic research this practice is usually not recommended as specifications should ideally be backed by theory, but these criteria are still useful to compare competing models when there is uncertainty.

distribution for the data, this otherwise very commonly used test cannot be performed. Still, with these statistics it is possible to compute what is known as the pseudo-R^2 of the model. This value is defined as follows:

$$R_D^2 = 1 - \frac{D(x;\hat{\mu})}{D(x;\bar{x})} \tag{2.11}$$

The pseudo-R^2 shows the relative reduction in deviance when the covariates are included in the model. Due to its definition as a rate, it can be used even when the data does not follow the likelihood used in the definition. Its name is due to the fact that it shares several characteristics with the linear R^2: its value is comprised between 0 and 1 and it can only increase as the number of covariates increases. It shares the same drawback than the linear R^2 as including random variables might increase its value and entice over-parametrised specifications. However, in this work I will not do a model selection based on any of this measures, so there is no risk of incurring in this bad praxis. Furthermore, in the paper Heinzl and Mittlböck (2003), the authors carry a MonteCarlo simulation study measuring the performance of this and some adjusted pseudo-R^2 coefficients when working with Poisson models in the case of over or underdispersion, and find that although some modified versions of the R^2 perform generally better, the regular R^2 works well when the number of observations is high, as in this work.

It could be argued that in the same manner that the use of the information criteria is deemed to be inadequate for models based of pseudo-distributions, so should be the use of the deviance, which are also based on the likelihood function and this is not in consonance with the pseudo-models philosophy. This is a natural concern, and the selection of one origin distribution or another, which can be seen as too arbitrary, will certainly lead to different deviances in each case. Still, it is preferred to work with deviances in these cases as information criteria are absolute measures which inform on how likely a model is given an underlying distribution, which is meaningless in this case, while deviances are relative measures that explain how much less informative a model is compared to the most parametrised model. Although this latter comparison depends on the underlying distribution that is chosen, it should change more smoothly when other distributions are considered, and in any case it is possible to attain a deviance equal to zero when the model perfectly predicts the data. Besides, the pseudo R^2 based on the deviances is a relative measure of goodness-of-fit that has quite useful characteristics. The only important point that should be made is that all these measures should be read from what they are: arbitrary statistics that depend on the underlying distribution that has been chosen and which can be more or less robust across changes in the underlying distribution chosen. Still, the coefficients obtained from models based on pseudo-distributions are consistent in many cases where certain conditions are met (see White (1982)). And that once that the researcher decides not to assume any underlying distribution, inertia from common practice should not lead to the use of (in this case) meaningless goodness-of-fit tests such as the deviance χ^2 test.

Once that the specification of the Poisson Pseudo-Maximum Likelihood model has been presented and the correct practice to work with it has been clarified the specification should

be clear. Assuming that the average conditional tourism flow is always positive and so are the explanatory variables, equation 2.1 can also be expressed applying both logarithms and an exponential as:

$$E(X_{ijt}|Z) = \exp\Big\{Z_{ijt}^{log}\beta\Big\} \tag{2.12}$$

where Z_{ijt}^{log} is the vector of values of all chosen explanatory variables after applying logarithm to them. Note that this specification is well defined even when X_{ijt} includes zeros, so it can be uses even in that case by extension. This is exactly the structure of equation 2.5 and the one that is required for the PPML estimator to be consistent. Also note that now the dependent variable is the untransformed number of travels, so the bias resulting from taking logarithm as identified by Silva and Tenreyro (2006) in the OLS specification disappears.

Equation 2.12 is also useful to understand the interpretation of the coefficients resulting from a PPML regression. Although the dependent variable of the regression is not transformed unlike in the case of the log-log specification, the interpretation of the coefficients can still be very similar to the normal case. Applying logarithms to both sides of the equation it is possible to see that the coefficients represent elasticities of the conditional expectation of travels on the the explanatory variables if a logarithm has been previously applied to them, or the semi-elasticities in the case that it has not[6]. Note, however, that the interpretation does somewhat change when elasticities are not equal across regions: an OLS regression estimates the average elasticity of a variable, while the PPML regression estimates the elasticity of the average coefficient. This subject is treated in depth in Tyazhelnikov and Zhou (2020)). Still, saving the nuance in the meaning of the coefficients, both methods are easily comparable, and I follow Silva and Tenreyro (2006) in assuming that the differences between them are due to the bias of the OLS specification.

2.3 The determinants

In this work I will base on Morley et al. (2014) to include some explanatory variables backed by the theory. Following the paper, where a gravity-model formulation is derived from consumer theory, people travel so as to benefit from the site qualities of their destination, proxied by a vector of variables ZD_{jt}, which compete against a vector of other goods, Q_{it}, from the perspective of the region of origin. The authors also allow for some influence of the origin characteristics, proxied by some variables ZO_{it}. Assuming individual's rationality, the number of travels by each individual to each location can be computed by solving their maximisation program:

$$\begin{aligned} max\ U_{ijt} &= f(N_{ijt}, Q_{it}, ZO_{it}, ZD_{jt}) \\ \text{subject to:}\ \ & \pi_{ijt}N_{ijt} + p_{it}Q_{it} = M_{it},\ N_{it} \geq 0,\ Q_{it} \geq 0 \end{aligned} \tag{2.13}$$

[6]The semielasticity of a function $f(x_1, ..., x_n)$ with respect to a variable x_i is defined as the marginal percentage change of f derived from a unit change in x_i. Algebraically: $S_i f(\vec{x}) = (\partial f(\vec{x})/\partial x_i) f(\vec{x})^{-1} = \partial \ln f(\vec{x})/\partial x_i$.

Overlooking concerns about aggregating demand, we obtain that tourism demand is a function of the price of the competing goods in the region of origin, p_{it}, the price of traveling, π_{ijt}, the income in region of origin i, and an origin and destination set of unilateral variables. The multiplicative formulation results from assuming a power model modeling tourism demand and computing the solution of the maximisation program.

This short derivation makes clear the theoretical motivation for the inclusion of some specific variables in the model. These are:

- The price vector of the consumption goods in the region of origin, p_i. In this work this variable will be proxied by origin region CPI. It is surprising how very rarely this variable is included in models, although its motivation from the theoretical derivation is clear.

- The income level in region i. The necessity of this variable would explain the inclusion of origin GDP per capita, which could be included in vector ZO_{it}.

- The average cost of visiting destination j from i. Here are involved three main types of variables. The first one would be transportation costs, ideally flight or train fares or an estimation of road-trips costs. Since this data is hard to find, distance between regions, in one of its several possible definitions, is used. The second one is price of goods of the destination. The higher prices in the destination, the more costly the travel and benefits from it will be. Again, this can be proxied by CPI in the destination. Finally, Morley et al. also include here other "psycho-geographical" variables such us speaking the same language or religion, having been in a colonial relationship, sharing the same border, etc. Out of them, the variables that could have an effect in the case of the Italian regions is sharing the same border and travelling from or to an island, whose effects will be studied.

- The authors also specifically suggest the inclusion of destination's GDP per capita as it can be interpreted as a destination quality indicator, proxying other factors such as security and health at the destination.

- Finally, as these are factors explaining average individual's demand to travel, population of origin has to enter the equation when the dependent variable is absolute travel. The aggregation concerns can, at least to some extent, be appeased by the fact that it is allowed to enter with an elasticity different to one.

Focusing on the theoretical derivation of the gravity equation has already allowed for the choice of some explanatory variables. Still, the origin and destination qualities vectors, ZO_{it} and ZD_{jt}, are far from being strictly defined. Similarly to Massidda and Etzo (2012), I have been interested in measuring the impact of the culture offer and the crime rate as an attraction and repulsion respectively for traveling to a specific region. Another variable of interest that they include is the number of international trips made by the residents of the region of origin. From the consumer-theory derivation it is clear that not accounting

for the role of other non-Italian destinations can be a source of bias, as they are directly competing with national destinations. For example, the fact that international travels are likely to be more common from northern regions which share a border with other European countries, can lead to a downwards bias in the estimation of the elasticity of GDP per capita[7]. Since I could not find data on international travels I have accounted for this possible bias including a dummy variable equal to one if the region shares a border with an European country and equal to zero in the opposite case. Finally, I also include population of the destination since it can proxy tourism offer and does not suffer from reverse causality like number of accommodations. It also correlates with other characteristics such as resistance to over-touristification and number of attractions not included in the "culture" variable.

To test Linder's hypothesis I include a variable that proxies the economic distance between every pair of regions, having for a given flow a value equal to the absolute value of the difference of the origin and destination region's GPD per capita. Linder's hypothesis is relevant for the case of trade off services and in particular for tourism. A reformulation for the latter would be that, as regions with similar incomes per capita share similar factors's prices, they specialise in similar services for consumers, and rich regions would have more high-quality services that attract mainly tourists from other rich regions. Therefore, as the economic distance increases, Linder's hypothesis would predict that flows between regions decreases, implying a negative coefficient.

Once that the specification includes all bilateral variables of interest and accounts for the role of international destinations as competitors, there remains the effect of competition of every region against each other to fully include the multilateral resistance effect. Note that this issue, which I have briefly described in the general framework of the trade gravity model, has the same importance when addressing tourism flows. The different approaches taken by the authors working on tourism differ in the same way that they do on general trade. I will base on Durbarry (2008) and use price to construct this variable in the following manner:

$$\ln MR_{ijt} = \ln \frac{\sum_k x_{ikt} PCI_k}{PCI_j}, \tag{2.14}$$

where x_{ikt} is a weight and is computed the fraction of the number of travels from origin i to every possible destination k excluding j over the total travels from i at time t. The higher the multilateral resistance term, the more attractive destination j is compared to the competition. Therefore, a positive coefficient

[7]Out of the 21 Italian regions, taking Trentino and South Tyrol as separate entities, those which share a border with another European country are also in the top eleven regions ordered by GDP per capita. Or with a different perspective, only Trentino, Lazio and Toscany have a higher GDP per capita than the Italian average and do not share a border with another European country

Chapter 3

Data and empirical results

3.1 The data

The empirical study will use the annual flows among all Italian regions for the time span between 2009 and 2019, with both years comprised. It was obtained from South Tyrol and the Autonomous Province of Trento are both included as separate regions summing up to 21 possible entities. Our dependant variable is the number of arrivals to every one of the 21 regions from the 20 possible origins (intra-region tourism is discarded) during all considered years, therefore summing up to 5040 flows (the panel dimensions are $N = 420$ and $T = 12$). It was obtained from the census "Movimento dei clienti negli esercizi ricettivi" conducted by the ISTAT (the Italian National Institute of Statistics). There is a flow value equal to zero corresponding to the number of tourists going from Bolzano - Bozen to Marche in 2010. Whether this is the true value of the flow or it is the result of the limitations of the recollection of the data is out of the concern of this work, which takes the data provided by ISTAT without further considerations. Although it is unlikely that removing this single value from the set of 5040 flows will change the results, the already explained appeals of the PPML model are compelling enough as to implement it, allowing for the inclusion of the null flow in the specification. In appendix 5.1 I show an image of the graph of aggregated travels by pair of regions in 2019.

The explanatory variables can be divided in nature into quantitative and qualitative variables, and also among the classes of variables defined at the origin, at the destination and linking both origin and destination.

The class of variables defined at origin of the flow include population ($population\_or_{it}$) and GDP per capita of the region of origin (GDP_or_{it}), defined as nominal GDP divided by total population. The necessity of using nominal GDP is clear from the theoretical derivation and the inclusion of of origin CPI (CPI_or_{it}). The rest of the variables that characterise the region of origin are time-invariant dummies: sharing a border with another European country (eu_border_i) and being an island ($island\_or_i$). These two dummies are expected to have a negative sign. Sharing a border with an European country should diverge part of the domestic tourism flow versus other European countries that become comparatively more

attractive compared to domestic tourism than for the rest of the regions. Regarding the islands, the associated isolation results in comparatively higher transport costs for domestic tourism, whose attraction compared to international tourism should become lower than for the continental countries.

The next groups includes the variables characterising the destination. Population, GDP per capita and price levels of the destination are defined equivalently as in the case of origin. Besides, the effects of culture offer ($culture_j$) as an attraction and crime rates ($crime_{jt}$) as a repulsion are also studied. While the crime variable is time-variant, I could not find consistent data on culture sites at a region level for all the years studied. Thus, culture is a time-invariant variable that reflects the number of cultural attractions of the region of destination in 2015. This variable is unlikely to suffer steep changes in the time span considered so the lack of information should not be too significant. Finally, the only dummy of importance regarding the destination is being an island: $island\_dest_j$. Although the associated travel costs should disincentivise travels to the islands, the specific attractions of and island might compensate this drawback making them more competitive compared to other destinations, so making a prediction in this case may not be so easy.

The last class refers to the bilateral variables mainly includes $distance_{ij}$, which is defined as the road kilometers that separates the capitals[1] of regions[2]. i and j. The decision to use this definition and not simple cartographic distance comes from the fact that a great majority of Italian tourists use some sort of road transport to travel through Italy[3], and road distances can also proxy rail distances better than cartographic distance. The use of the capital of the regions is justified as it is a simple and methodical manner to set the node representing each region, and by the fact that the region capitals correspond to the most populated cities for almost all the regions[4]

The rest of the bilateral variables include Linder's variable for the economic distance between two regions ($linder_{ijt}$) and the Multilateral Resistance term (MR_{ijt}). Both are defined

[1]The capitals of the regions are the regions' *Capoluoghi* (plural of *Capoluogo*), where the regional council is located. Note that in Italian the usage of the word *capitale* is restricted to the capital of a country, so Rome would be the *capitale* of Italy but Milan is the *capoluogo* of Lombardy. Rome is also the *capoluogo* of its region: Lazio.

[2]For the case of the islands, distances have been defined as the shortest driving plus ferry distance between their capitals and every other. The distance has been taken from Google Maps since the ISTAT does not have data for it. I have checked that the distance given by Google Maps is similar to that given by ISTAT for the rest of the distances so that the definition is the most consistent possible

[3]In 2019 for example, the combination of automobiles, buses and campers accounted for 81% of total travels, compared to less than 7% of airplanes (Data obtained from *ISTAT. Viaggi e loro caratteristiche: Mezzo di transporto e destinazione*).

[4]They are: Veneto (whose most populated city is Verona with 257 748 inhabitants, approximately two thousand more than Venice in January 2021), Abruzzo (whose most populated city is Pescara with 119 327 inhabitants while its capital L'Aquila has 69 996) and Calabria (with 173 367 inhabitants in Reggio Calabria while the capital, Catanzaro, has 86 606). In the case of Veneto the population difference is very small to really be a concern of the population rule, while Abruzzo is small enough in its extension as to have a very small effect on the distance variable. The only case that could raise a concern is Calabria, although being one case out of twenty one and the driving distance between Catanzaro and Reggio not surpassing the 160km it can be safely assumed that using population to define the nodes would not significantly change the results.

as in section 2.3 on tourism determinants. According to Linder's hypothesis the coefficient associated to Linder's variable should be negative as the more different the countries are the less trade should take place, while the higher the multilateral resistance term the more comparatively accessible is destination j and therefore more tourism should be expected to happen, thus yielding a positive coefficient. The only dummy variable in this group is the one expressing if regions i and j share the same border or not ($border_{ij}$), whose sign is also expected to be positive. Adding year effects, which do not depend on any region, all variables studied are included. Table 3.1 is a compilation of all the variables along with a short description, their main characteristics and their source. Finally, the main descriptive statistics of the quantitative variables are shown on table3.2.

Table 3.1: Description of the explanatory variables

Variable	Definition	Nature	Region	Time varying	Source
arrivals	Number of arrivals registered by region of origin and residence and year	Q	B	Yes	ISTAT: Movimento dei clienti negli esercizi ricettivi.
distance	Driving distance (in km) among the capitals of the origin and destination regions	Q	B	No	ISTAT: Matrice delle distanze and GoogleMaps.
GDP_pc	Nominal GDP (as in December of the corresponding year) divided by total population	Q	OD	Yes	ISTAT: Prodotto Interno Lordo lato produzione.
population	Population by region counted the 1st of January	Q	OD	Yes	ISTAT: Popolazione residente al 1º di gennaio.
CPI	Consumer Price Index for the entire collectivity taking 1998 as the reference value. General index	Q	OD	Yes	ISTAT: Nic. Medie annuale. Classificazione Ecicop (3 cifre)
border	Signals if the regions share a border	D	B	No	Defined ex profeso
eu_border	Signals if the reggion shares a border with a European country	D	O	No	Defined ex profeso
island	Signals if the region is an island	D	OD	No	Defined ex profeso
culture	Number of museums and other cultural institutions (galleries, archaeological sites, monuments and other places open to the public) in 2015	Q	D	No	ISTAT: Musei ed istituzioni similari.
crime	Total number of crimes reported by the police regardless of whether the identity of the offender is known by 100 000 inhabitants	Q	D	Yes	ISTAT: Delitti denunciati dalle forze di polizia all'autorità giudizaria
linder	Absolute value of the difference between the GDP per capita between the region of origin and that of destination	Q	B	Yes	Defined ex profeso
MR	Multilateral resistance term as defined in Chapter 2	Q	B	Yes	Defined ex profeso
year	Year effects	D			

Note: The name of the variable corresponds to the name given in the code except for the origin/destination marker. Nature refers to the type of variable, whether quantitative (Q) or dummy (D). Region expresses the region of definition of the variable. B stands for Bilateral (the variable is distinctly defined for each Origin-Destination pair), O and D mean that the value of the variable for a flow only depends on the region of origin or destination of the flow respectively. OD means that the variable enters twice into the equation, accountig for the value in the origin and the destination separately and are called $< Variable > \_or$ and $< Variable > \_dest$.

Table 3.2: Main descriptive statistics of the explanatory variables

	mean	sd	median	min	max	range	skew	kurtosis	se
arrivals	104428.48	173262.36	41114.50	0.00	2182620.00	2182620.00	4.54	32.71	2440.56
distance	615159	340137	558458	58537	1588000	1529463	0.67	-0.05	4791.14
GDP_pc	0.03	0.01	0.03	0.02	0.05	0.03	0.25	-0.86	0.01
population	2853452	2444421	1650793	125653	10010833	9885180	1.04	0.73	34432
crime	8.24	0.24	8.21	7.70	8.72	1.02	0.17	-0.76	0.01
culture	5.25	0.72	5.38	3.74	6.31	2.57	-0.59	-0.45	0.01
CPI	135.80	6.62	136.50	121.60	155.20	33.60	-0.07	-0.44	0.09
MR	1.07	0.04	1.08	0.98	1.14	0.16	-0.61	-0.99	0.001

Note: Table computed using R's function `describe` from the `psych` package.

3.2 Results

To run the model I have used the library `Gravity` of R, which includes a specific `ppml` function which runs a PPML model, although the default `glm` function could have been used as well obtaining the same results. I have also run an OLS regression on the transformed dependent variable $\ln X_{ijt} + 1$ to compare, and in both cases I have checked the results both with and without time effects. The preferred specification in case of discrepancies corresponds to the PPML with time effects. Details on the values and significance of the estimated elasticities are shown on table 3.2.

The results are generally in accordance to what it was expected. Changing the specification from an OLS with dependent variable $\ln X_{ijt} + 1$ to a PPML where the dependent variable is not transformed, leads to fairly different estimated coefficients for most variables. Including or not year effects leads to significant changes, especially in some time-varying variables such as GDP_pc_or or the CPI variables. This could be due to a previous partial assimilation of the influence of omitted time trends by these variables due to its monotonic changing nature. As it was expected, the deviance of the PPML models is very high, implying that the data does not follow a Poisson distribution, but the pseudo-R^2 shows a good fitting of the model both with and without time effects. Regarding the coefficients, all the classic gravity model variables are significant and have the expected sign. Still, other variables have coefficients which are contrary in sign to what was expected or are not significant, deserving some remarks to analyse the reason.

For those defined at the origin, GDP per capita shows significant positive values for both versions of the PPML regression, although it is somewhat higher for the case of time effects. The result is comparatively smaller than that obtained by Massidda and Etzo (2012) of 1.4, but quite similar to the one obtained by the equivalent study on Spanish domestic tourism done in Priego et al. (2015), implying that domestic tourism in both countries is not a luxury good[5]. This could be the result of the difference in the statistical models used, although the fact that Massidda and Etzo (2012) includes a proxy for international trips, or that it does

[5]Luxury goods are usually defined as goods for which its demand increases more than proportionally as income increases. Since, as discussed in the work, domestic tourism is a substitute of the most expensive luxury tourism, which I could not proxy, the low GDP elasticity could be the result of the omitted variable bias.

Table 3.3: Means and Standard Deviations of Scores on Baseline Measures

Estimator:	OLS	OLS (Time eff.)	PPML	PPML (Time eff.)
Dependent variable:	$\ln X_{ijt} + 1$	$\ln X_{ijt} + 1$	X_{ijt}	X_{ijt}
distance	-0.76***	-0.77***	-0.51***	-0.51***
	(0.02)	(0.02)	(0.01)	(0.01)
GDP_pc_or	0.10**	0.24***	0.38***	0.63***
	(0.04)	(0.04)	(0.03)	(0.03)
GDP_pc_dest	0.96***	1.03***	0.76***	0.85***
	(0.04)	(0.04)	(0.05)	(0.04)
population_or	1.03***	1.02***	1.04***	1.06***
	(0.01)	(0.01)	(0.01)	(0.01)
population_dest	0.45***	0.43***	0.31***	0.31***
	(0.01)	(0.01)	(0.01)	(0.01)
island_or	-0.43***	-0.38***	-0.23**	-0.18***
	(0.03)	(0.03)	(0.03)	(0.03)
island_dest	0.28***	0.34***	0.09**	0.13***
	(0.03)	(0.03)	(0.03)	(0.03)
eu_border	0.06**	0.02	0.04**	0.03*
	(0.02)	(0.02)	(0.01)	(0.01)
border	-0.10***	-0.13***	-0.05**	-0.06***
	(0.03)	(0.03)	(0.02)	(0.02)
crime	-4.01***	-3.35***	-2.35***	-2.16***
	(0.43)	(0.48)	(0.33)	(0.37)
culture	2.05***	1.94***	1.01***	1.07***
	(0.11)	(0.11)	(0.09)	(0.09)
linder	0.06***	0.05***	-0.01	-0.02
	(0.01)	(0.01)	(0.01)	(0.01)
CPI_or	-4.55***	-1.10**	-0.24	4.13**
	(0.25)	(0.33)	(0.24)	(0.37)
CPI_dest	5.94***	10.03***	1.44***	4.34***
	(0.26)	(0.36)	(0.23)	(0.29)
MR	4.09***	4.63***	3.22***	3.40***
	(0.12)	(0.12)	(0.04)	(0.04)
Year effects	No	Yes	No	Yes
Null deviance:			846039967	846039967
Residual deviance:			83642617	79994608
R^2 and pseudo R^2	0.900	0.902	0.901	0.905

Note: The OLS estimation has been performed using the `lm` command in R, while the PPML estimator has been applied using the `ppml` function from the package `Gravity` also in R. The estimated coefficients are shown in the same level of the variable name, while each estimated standard deviation is shown under it. The deviance and R^2 values are obtained applying the function `summary` to the respective models and the pseudo-R^2 was computed using the deviances. Stars denote p-values as follows: $^*p < 0.05$; $^{**}p < 0.01$, $^{***}p < 0.001$.

not include an aggregate "mass" variable (e.g. population or GDP) may also play a role. For its part, the elasticity of population revolves around 1, so an increase in one percent of the population results in an increase of one percent in tourism flows, which is exactly what it was expected from the theoretical derivation and points to the validity of the model. Note that some authors working with GDP obtain substantially smaller estimates, both for trade (as was the case of Silva and Tenreyro (2006)) and for tourism (as is in the case of demand for tourism in Greece studied in Chasapopoulos et al. (2014)). Still, most papers do obtain elasticities close to 1, for example see again Priego et al. (2015).

Other variables, *island_or* and CPI_or, are less conclusive. The first has the expected sign but it is not significant in the preferred specification, and the second has mixed results: from being negative or not significant, to somewhat significant and positive (as expected) in the preferred specification. Finally, *eu_border* seems to be somewhat significant but with a positive sign, so belonging to a region with an European border actually increases the expected flow to other Italian regions. This surprising result may be due to some omitted variable bias. These regions, which are also some of the richest in Italy, may have other characteristics not accounted for in GDP per capita that foster Italian tourism demand.

Moving on to the variables defined at the destination, GDP per capita and population have again the expected signs and a high significance level. Surprisingly, GDP per capita seems to play a bigger role at the destination than at the origin (maybe due to the bias discussed before). On the other hand, *population_dest* has (as expected) a positive significant elasticity which is a bit less than half of the origin population elasticity. For its part, both crime and culture have very consistently negative and positive elasticities respectively, showing that they do repel and attract tourism in the manner that was expected. On the contrary, CPI_dest does show a consistently significant effect, but it is opposite in sign to what it was expected. Again, it may be possible that the prices at destination correlate with hidden variables that are not accounted for with GDP per capita. This possibility will be discussed in a moment. Finally, the island destination dummy shows a significantly positive effect, which might be surprising due to the associated transport restrictions, although the own characteristics of being an island, for example kilometres of coast which is not accounted for, might explain it.

Finishing with the bilateral variables, distance shows a fairly more negative elasticity, -0.51, compared to the value of -0.2 that some authors (e.g.Massidda and Etzo) take as a reference for international tourism following Khadaroo and Seetanah (2008). Yet, it remains quite lower in absolute value than the one of -0.9 found by Priego et al. for in Spain or even the one of -0.7 that Culiuc estimated for international flows. In any case, the fact that distance is a stronger barrier for domestic tourism than for the international one should not be surprising, if only because domestic tourism is mainly performed by road trips, while the international one by airplane travels[6]. Another argument in favour would be the role of

[6]My reasoning is that airplane travel implies costs much less proportional to distance than road costs, being the first a very complex variable depending on many market factors while the second mainly depends on fuel consumption and possible motorway tolls. Also, travel time is almost proportional to road distance for car trips, while this is much less clear for plane travels, which also include transportation to the airport,

exotism in international travel as studied by Mayo et al. (1988). Surprisingly, Massidda and Etzo (2012) find a elasticity of -0.07 for the Italian domestic tourism flows, concluding that in fact domestic travel has an elasticity with respect to distance smaller than international tourism.

The effect of the other bilateral variables is more abstruse. Sharing the same border seems to have a negative effect, although quite small compared to that of distance, and contrary to what most papers find: in most cases is a positive sign (including Priego et al. (2015) for domestic tourism and Culiuc, 2014 for international tourism); and sometimes a non-significant coefficient (for example Silva and Tenreyro (2006) for trade). Still, there is the possibility that sharing a border might have a downwards effect in arrivals, as an important share of visits would maybe be one-day trips, and actual hotel stays are proportionally higher in other regions once distance is accounted for. On the other hand, Linder's variable fails to show influence as a determinant. This could be considered a point in favour to the gravity model compared to Linder's theory, based in different premises. Finally, the multilateral resistance factor shows a strong effect in the preferred specification and with the expected sign. The fact, however, that price in the destination does not have a negative effect may raise concerns on the interpretation of this result. If high prices in the destination do not deter but actually increase tourism, high prices in competing zones compared to the destination should decrease and not increase the flow of tourism.

Finally, I have run some tests checking the robustness of the model. First, I have run an overdispersion test to see if the use of the pseudo-poisson model is justified or if instead a Poisson regression could have been suitable. I have applied the test proposed in Cameron and Trivedi (1990), and which is easily performed in R thanks to the package `AER`, which includes it in the function `dispersiontest`[7]. The statistics of the test for equidispersion against and alternative hypothesis of overdispersion had an associated p value of the lowest number possible in R, therefore the null hypothesis is rejected along the possibility of the implementation of a Poisson model.

More information on this matter can be obtained from the plots of the residuals. In figure 5.2, I have plotted the deviance residuals, which are defined for each raw residual as the root square of its contribution to the residual deviance (defined in equation 2.9), multiplied by

time spent in controls, etc; which are fixed or not very correlated with distance. Thus, distance should have a clearly lower effect in international tourism than in the domestic one, as found in the work.

[7]The functionning of the test is fairly straightforward. The variance is formulated in the following manner: $V(X_{ijt}) = \hat{X}_{ijt} + c * f(\hat{y})$, where $f()$ is a monotonic increasing function and c is a parameter indicating overdispersion if $c > 0$ or underdispersion if $c < 0$. Then, the null hypothesis is $H_0 : c = 0$ and the alternative $H_1 : c \neq 0$. The test statistic is a $\mathbf{t}$ statistic which is asymptotically standard normal under the null. There are also one-side versions of the test to specifically test for over or underdispersion as the alternative hypothesis.

the sign of the raw residual[8]. Deviance residuals are more robust to the inclusion of zero values and are somewhat preferable than Pearson residuals for the case of GLM (see Dunn and Smyth (2018)). From the figures it is possible to observe the overdispersion of the data along with some error clustering that was not perfectly accounted for[9] (see the queues at the right part of the plot). In figure 5.3 I also present a plot of Cook's distances and a residuals vs. leverage plot. These plots are useful to know if excluding certain extreme observations could lead in steep changes in the estimation. Cook's distances remain very low, with some peaks corresponding to the flows between Campania and Molise but not important enough to be considered outliers. On the other hand, no outliers are present for very high values of the leverage. Therefore, there should not be strong concerns of atypical influencial values, although some authors caution on the interpretation of these graphs for generalised linear regressions. To see more on diagnostics for generalised linear regressions see Dunn and Smyth (2018).

Another concern regarding the model can refer to the selection of the explanatory variables and the changes resulting from their exclusion. Since the effect of prices on tourism is probably the least clear (both a high CPI in the origin and in the destination seem to have a similar positive effect), I also have run an alternative model where origin and destination CPI enter as a single variable, called CPI_rate. This variable is defined as the rate of prices at destination divided by prices at origin. This specification is very common in other works using the gravity model (including Massidda and Etzo (2012) or Culiuc (2014)), and allows for a more simple entrance of prices in the model. The results of the regression are shown in the appendix, in table 5.2. The pseudo R^2 of the new model is very close to the old one, losing very little information from this change. Surprisingly, the positive coefficient implies that higher prices at the destination compared to the origin attract more instead of less tourism. However, other works do find negative elasticities for this variables, for example, around -0.3 in Culiuc (2014) or -8.97 in Massidda and Etzo (2012) (who use the same data and definition).

As one of the differences of this work with respect to others is the inclusion of the MR effect, which is defined by using CPI in the destination, I ran a third model excluding the MR factor and maintaining the relative CPI variable. The results are shown table 5.3. Even if the MR term and the relative CPI do not have a strong correlation[10], both PPML regressions (with and without time effects) now yield negative and non significant coefficients once MR has been left out, although in this case the pseudo R^2 value significantly decreases to 0.837. These results may suggest a possible bias arising from the inclusion of

[8]For the specific case of a Poisson regression the deviance residual of observation i equals to:

$$d_{ijt} = sign(X_{ijt} - \hat{X}_{ijt})\sqrt{2[X_{ijt} \ln \frac{X_{ijt}}{\hat{X}_{ijt}} - X_{ijt}\hat{X}_{ijt}]}$$

[9]This is probably the result of not including regions fixed effects and could have lead to some bias in the estimations.

[10]They have a correlation of -0.14 before applying logarithms and a correlation of 0.26 after applying them.

the multilateral resistance term in the model, which being negatively correlated and having a positive coefficient, could bias upwards the coefficient of the relative price. Next chapter will help to conclude whether these coefficients are robust or not when different variables are excluded of the model.

In conclusion, it has been possible to corroborate the effect of the classic determinants of tourism for the case of the domestic Italian flows. In particular, distance, population and GDP per capita, but also culture or crime. Furthermore, some elasticities have been shown to be close to the results found by other authors and in the case of discrepancies it has sometimes been possible to trace the cause of the deviations. However, a big question mark remains on the actual role of the less significant or more volatile variables such as Linder's and having a border with another European coutnry, or even those regarding prices and the MR term, whose interpretations are not totally clear. Besides, working with a model based on a pseudo-distribution reduces the amount of statistical tools that can be used to check the robustness of the specification, or at least makes their interpretation trickier. Here is where the Bayesian Model Averaging technique, which, with all due caution, is compatible with pseudo-distributions and takes a less strict approach than other model selection algorithms, can be helpful to double check the robustness of the estimated coefficients.

Chapter 4

Bayesian Model Averaging

In the second chapter some of the most commonly studied tourism determinants have been discussed and proposed, first from a purely theoretical derivation and then getting inspiration from other empirical papers. Then, in the third chapter a Poisson Maximum-Likelihood model has been proposed for the estimation of their elasticities, which have been compiled in table 3.2. From their statistical significances it is possible to draw conclusions on their actual relevance regarding the study of (at least Italian) domestic tourism flows, and it would have been possible to finish the work with a set of statements such as "income is a clearly positive determinant of domestic tourism", or "Linder's hypothesis does not hold in this specific case".

This common practice suffers from the limitation of only focusing on one possible model to draw conclusions. The amount of models that have been applied by different researchers to model tourism flows might raise some doubts on the validity of the model presented in this work. Although the usefulness of applying a PPML model to the data has been discussed, hopefully extensively enough so that it can be accepted as good practice, the selection of the variables has been influenced by different sources and it is unclear how a different specification could affect the conclusions. Due to correlation, the inclusion and exclusion of a certain variable can severely affect the estimates of the elasticities of the rest of the variables. Although the correlation matrix of the variables has been studied to make sure that no pair of variables is close to show collinearity[1], some correlation still exists, and when the number of variables is as high as in this case it is very likely that taking different combinations of the covariates results in different conclusions for each model.

In order to mitigate this uncertainty, a Bayesian Model Averaging procedure will be implemented. Starting with the initial model where all the presented explanatory variables are included, a classic frequentist approach to model selection would usually end up with a subset of variables that maximises some information criterion (namely Akaike's or BIC). On

[1] A very high correlation among a set of variables makes their estimates very sensitive to the inclusion or exclusion of one or a subset of these variables. If the correlation of a pair of variables is very high, one of them is already proxying of the other so it must be excluded. When there is perfect collinearity the estimation simply cannot be computed.

the contrary, a model averaging strategy consists of taking into account all candidate models and then draw conclusions by giving different weights to them. These weights correspond to the perceived probability that a given model is the correct one. Once these weights are endowed, it is possible to compute a new coefficient for each variable as the weighted average of the models estimates of the coefficients where the weights are the model probabilities. Furthermore, a distribution for each coefficient can be obtained, and also its probability of being different from zero. The interest of model averaging is that no model is unilaterally chosen to be the correct one, but uncertainty is instead modeled and used to draw conclusions. Although there also exists a frequentist version to model average (as defended in Moral-Benito (2011)), model averaging procedures have come from a bayesian context and imply the specification of priors for both the models considered and the coefficients to be estimated.

4.1 Bayesian statistics

The classic philosophy of statistics is based on the assumption that with enough observations it is possible to obtain a more accurate depiction of a the data generating process under a phenomenon of study. This corresponds with the frequentist approach, which is backed by theorems like the Central Limit Theorem and the Laws of Large numbers, and which allows for an (at least asymptotically) objective estimation of the parameters of a model.

The bayesian paradigm takes a different perspective. Instead of hoping that the number of observations will be high enough for the asymptotic theorems to be good enough approximations, a bayesian statistician in a parametric context already starts with some assumptions on the coefficients that they want to estimate. After they are confronted with the data, they modify their previous believes incorporating the new information into their model. For example, in order to estimate a regression model, the researcher may start with an assumed distribution for the vector of parameters to be estimated, $Pr(\beta)$. Then, given a dataset D, composed of a set of independent variables and a dependent variable, the distribution of the parameters can be updated following Bayes' theorem:

$$Pr(\beta|D) = \frac{Pr(D|\beta)Pr(\beta)}{Pr(D)}. \tag{4.1}$$

Each term of this equation receives a specific name, referring to its role and nature in a bayesian framework. $Pr(\beta)$ is the *prior probability*, reflecting the previous believes of the researcher on the parameters that they want to estimate. $Pr(\beta|D)$ is the *posterior probability*, and corresponds to the modification of the prior to adapt to the evidence. $Pr(D|\beta)$ is called the *likelihood*, indicating the compatibility of the data with a hypothesis. Finally, $Pr(E)$ is usually termed the *marginal probability*. Although I have referred to a regression model for clarity, this philosophy can be applied to much more general contexts, in which we can talk about a hypothesis, H, whose probability can be affected by the data and for which we have a prior.

4.2 Theoretical framework of BMA

Bayesian Model Averaging is a special case of application of bayesian statistics in which not only the coefficients of a regression are initially endowed with a prior, but also the different models are treated as uncertain and given probabilities.

To implement the method, it is first of all necessary to establish the models $\{M_j\}_{j=1,...,k}$ that are deemed possible and associate a prior probability to all of them ($P(M_j)$ for $j = 1, ..., k$). In this case, the models considered will be all the possible models defined by taking all possible subsets of the set of explanatory variables $Z = \{Z_1, ..., Z_q\}$[2]. Thus, 2^q models can be estimated, each seeking to explain the same data y, and yielding their own set of coefficients $\{\hat{\beta}^1, ..., \hat{\beta}^q\}$. Considering a model M_j that does not include a certain variable Z_p is in this case tantamount to setting the corresponding value of the vector of coefficients equal to zero ($\beta_p^j = 0$). If we stack our data in D, which includes both the value of the dependent and the explanatory variables, then the law of total probabilities allows for the specification of a probabilistic distribution of the coefficients conditioned on the data. This is the posterior distribution of β, and is given by:

$$Pr(\beta|D) = \sum_{j=1}^{k} Pr(\beta|D, M_j)Pr(M_j|D). \tag{4.2}$$

In 4.2 is possible to see here that the BMA posterior distribution is a weighted average of the posterior distributions of the vector of coefficients β under each of the models, where the weights are the posterior probabilities of the models. They are given by:

$$Pr(M_j|D) = \frac{Pr(D|M_j)Pr(M_j)}{Pr(D)} = \frac{Pr(D|M_j)Pr(M_j)}{\sum_{l=1}^{k} Pr(D|M_l)Pr(M_l)}. \tag{4.3}$$

In order to work with expression 4.3 it is necessary to be able to compute the likelihood of the data when it is conditioned on each model. This is obtained by integrating the likelihood of the data over the unknown coefficients[3]:

$$Pr(D|M_j) = \int Pr(D|\beta, M_j)Pr(\beta|M_j)d\beta \tag{4.4}$$

The first factor, the likelihood of the data given the vector of coefficients and the model has to be specified by the researcher. In the case of a model based on a pseudo distribution the most natural election would be to use the origin distribution, in this case a Poisson. The second one is just the coefficients prior, which must also be chosen by the researcher.

[2]The probabilistic space of models is simply equal to the power set of Z on which it is defined a probability measure, where Z is understood the set whose elements are the variables $Z_1, ..., Z_q$.

[3]Remember that given three random variables X_1, X_2 and X_3 with joint density $f(X_1, X_2, X_3)$, the conditional density function of the first random variable over the others is defined as: $f(X_1|X_2, X_3) := f(X_1, X_2, X_3)/f(X_2, X_3)$. Therefore: $\int f(x_1|x_2, x_3)f(x_2|x_3)dx_2 = \int f(x_1, x_2, x_3)/f(x_3)dx_3 = f(x_1|x_2)$

Therefore, this integral is theoretically computable once the assumptions are clear. However, the high dimensionality of the integral can make the estimation by numeric methods too hard. And although in the linear regression case where the errors are assumed to be normal it has an analytical expression, this is not generally the case.

There are several approaches that are followed in order to cope with this integral. Many researches use MonteCarlo methods to reduce the computational complexity of its resolution. In this work I use the package of R `BMA`, which solves the challenge by approximating the integral by using BIC (see Raftery et al. (2005)). The approximation is as follows:

$$2 \ln Pr(D|M_j) \approx 2 \ln Pr(D|\hat{\beta}^j) - d_j \ln n = -BIC_j, \tag{4.5}$$

where d_j is defined as the dimension of the vector of coefficients in the model and $\hat{\beta}_j$ is the Maximum Likelihood estimator of coefficient β under model j.

With this approximation, equation 4.3 becomes:

$$Pr(M_j|D) = \frac{\exp\{-BIC_j/2\}}{\sum_{i=1}^{k} \exp\{-BIC_i/2\}} \tag{4.6}$$

Once that the weights of equation 4.2 are computed, the posterior of the coefficients will be specified. If the expectation of both sides of equation 4.2 is computed and the conditional expectation of β over a model is approximated by the maximum likelihood estimator, it is possible to obtain a BMA estimate of β:

$$\hat{\beta}_{BMA} = E(\beta|D) = \sum_{j=1}^{k} E(\beta|D, M_j) Pr(M_j|D) \approx \sum_{j=1}^{k} \hat{\beta}^j Pr(M_j|D). \tag{4.7}$$

Note that if the interest is to compute an estimate of β, the described procedure leaves a very small space for the influence of the priors. Only the priors of the models play a role, and since they are almost universally taken as equiprobable, the bayesian nature of the method disappears and it could be possible to talk about a frequentist model averaging. Another value of interest is the Posterior Inclusion Probability (PIP) of a variable, which for a given variable p is only the sum of the posterior probabilities of the models which include that variable:

$$PIP(\beta_p) = Pr(\beta_p \neq 0|D) = \sum_{\beta_p \neq 0} Pr(M_j|D) \tag{4.8}$$

The PIP allows for a new conception of significance of a variable. It is common to establish a threshold for a variable's PIP to consider it significant. I will follow the common practice of considering that a variable is significant whenever its associated PIP is higher than 0.5 (Beck et al. (2017a)). In other words, whenever there is a higher probability of variable to be included than not it will be called significant.

Note that the model priors are defined so that all models are equiprobable, no variable could have a PIP equal to one, since the probabilities of the models which do not include that variable are strictly positive. In practice, however, it is very common that the most

likely models accumulate a very big part of the probability mass. These group of very likely models often share a big set of explanatory variables which are included in all of them, so some variables end up having extremely high PIP, which are rounded up to 1. Besides, many BMA algorithms do not actually run all possible models since it would be extremely heavy in computational terms, but instead use some sort of shortcut, often dropping all models without a given variable when the posterior probability goes down a lot after its removal (it is for example the case of the MC^3 algorithm described in Moral-Benito (2011)).

Note that the way the posterior expectation is defined also makes the estimates of the coefficients to be somewhat closer to zero than they would be in the case that the focus were limited to the models where those variables are included[4]. This latter value can also be of interest for the researcher and it is denoted as the conditional posterior mean (PMC). It is just the posterior mean conditioned on the variable being in the model:

$$PMC(\beta_p) = E(\beta_p|\ \beta_p \neq 0, D) = \frac{\sum_{j=1...k|\beta_p \neq 0} Pr(M_j|D)\hat{\beta}_p^j}{Pr(\beta_p \neq 0)} \tag{4.9}$$

Another estimate of interest can be the posterior standard deviation (PSD), which is equal to:

$$PSD(\beta_p) = \sqrt{\sum_{j=1}^{k} Pr(M_j|D)V(\beta_j|M_j, D) + \sum_{j=1}^{k} P(M_j|D)[\hat{\beta}_p^j - E(\beta_p|D, M_j)]^2}, \tag{4.10}$$

whereas the conditional posterior standard deviation (PSDC) is given by:

$$PSDC(\beta_p) = \sqrt{\frac{V(\beta_p|D) + E(\beta_p|D)^2}{Pr(\beta_p \neq 0)} - E(\beta_p|\beta_p \neq 0, D)^2}. \tag{4.11}$$

These are the key concepts that will be used to analyse the effect of the tourism determinants. For a more theoretical description on BMA I refer again to Moral-Benito (2011). Another matter of interest when applying a BMA can be the relationships between the variables, in particular as substitutes and complements, which can be analysed through the use of jointness measures. Although I will not treat this issue in this work, a theoretical explanation and application of the BMA perspective for the case when pairs of variables care studied together and in relation with each other can be seen in Beck et al. (2017b), where the author also applies a BMA to a gravity model specification.

It can be noted now that the BMA suffers from the same problems as most statistical tools when working with pseudo-distribution models: they are using a likelihood which is not assumed to be true. Without incurring in this discrepancy it would not possible to

[4]Remember from equation 4.2 that the BMA estimates of the coefficients are computed as their posterior expectations. This equals to the weighted mean of the coefficients estimates from the models where the weights are the model posterior probabilities. When a model does not include a variable in its specification the associated coefficient is just set to zero.

operate further than computing the coefficients and the variance of the model. The question is when using this likelihood can be too problematic. In the case of the BMA, the likelihood plays a role only in the computation of integral 4.4, in particular in the computation of the BIC by the package `BMA`, which plays a relative role in assigning the probabilities of each model with respect to the data. While changing the likelihood would change the posterior probabilities of the models, note that the PPML has been shown by Silva and Tenreyro (2006) to be consistent in its estimation of the elasticities for different cases of overdispersion, so the Poisson likelihood is already effective to fit models to the data, even if the variance is computed aside. Therefore, the BIC values should be good indications on which models are most likely according to the data in relative terms. Besides, the BMA is used a as a robustness method, and its balancing nature should assuage concerns of possible biases arising from its use in the context of a model based on a pseudo-distribution.

Regarding its implementation, there are, at least to my present knowledge, three packages available for R to perform BMA: the already cited `BMA`, the `BMS` (Bayesian Model Selection) package and the `BAS` (Bayesian Adaptive Sampling) package. The choice in favour of `BMA` was rather a necessity, since it is the only one which is compatible with generalised regressions. What is more, for the Poisson model it permits to choose whether to follow the equidispersion constraint or to estimate the robust variance separately, therefore being compatible with a pseudo-poisson specification. It has other interesting characteristics as well, namely its use of the BIC to approximate the integral 4.4, reducing a lot the computational requirements of the model. For a discussion on the R packages for BMA and more information on the `BMA` package that has been used see Amini and Parmeter (2011).

4.3 Results

I have run the bayesian model averaging algorithm from R package `BMA`. The results are shown on table 4.1. They include the variables names in the first column, along with their estimated PIP, expected value and standard deviation in the next ones. The names of the variables are in bold letters when they are considered significant ($PIP > 0.5$). The rest of the columns represent the estimates of the first five most probable models, along with their BIC and their posterior probability. Note how very small differences in the BIC value translate into very different posterior probabilities for the models.

The most likely model according to the BMA procedure, which accumulates a probability mass higher than 0.5, includes all variables except for the *eu_border*. This finding corroborate those found in the previous chapter, providing strong evidence against the inclusion of this previously significantly ambiguous variable, which has a PIP of only 15.5. This might be either an evidence against the hypothesis that international tourism is a substitute of domestic tourism, or is the result of a failure of the model at trying to proxy it with this variable. Another possibility could be that regions sharing a border with another European country actually travel more than those which do not. As it was already mentioned, this group of regions include many of the richest areas in Italy and there could be some correlation with an omitted variable. Still, it is hard to explain what is the nature of this omitted

Table 4.1: BMA results

	p!=0	EV	SD	model 1	model 2	model 3	model 4	model 5
distance	100.0	-0.517247	0.016656	-5.178e-01	-5.278e-01	-5.202e-01	-4.901e-01	-5.301e-01
GDP_pc_or	100.0	0.655134	0.033968	6.611e-01	6.557e-01	6.313e-01	6.614e-01	6.233e-01
GDP_pc_dest	100.0	0.844211	0.030314	8.485e-01	8.387e-01	8.450e-01	8.373e-01	8.352e-01
population_or	100.0	1.066591	0.008312	1.068e+00	1.066e+00	1.064e+00	1.067e+00	1.062e+00
population_dest	100.0	0.306722	0.010986	3.079e-01	3.054e-01	3.089e-01	3.015e-01	3.066e-01
island_or	100.0	-0.185623	0.030039	-1.849e-01	-1.841e-01	-1.837e-01	-1.940e-01	-1.828e-01
island_dest	99.6	0.129412	0.030365	1.312e-01	1.305e-01	1.306e-01	1.229e-01	1.299e-01
eu_border	15.5	0.004755	0.012298	.	.	2.982e-02	.	3.247e-02
border	88.6	-0.050466	0.023601	-5.634e-02	-5.894e-02	-5.662e-02	.	-5.913e-02
crime	100.0	-2.101190	0.373547	-2.115e+00	-2.111e+00	-2.160e+00	-1.934e+00	-2.162e+00
culture	100.0	1.072410	0.095050	1.065e+00	1.086e+00	1.065e+00	1.084e+00	1.086e+00
linder	76.9	-0.013810	0.008932	-1.794e-02	.	-1.721e-02	-1.890e-02	.
CPI_or	100.0	4.137500	0.372680	4.163e+00	4.069e+00	4.135e+00	4.170e+00	4.043e+00
CPI_dest	100.0	4.327922	0.294658	4.353e+00	4.274e+00	4.338e+00	4.319e+00	4.260e+00
MR	100.0	3.410229	0.042469	3.413e+00	3.411e+00	3.404e+00	3.404e+00	3.402e+00
nVar				15	14	16	14	15
BIC				-3.775e+04	-3.774e+04	-3.774e+04	-3.774e+04	-3.774e+04
post prob				0.575	0.169	0.094	0.083	0.044

Note: Bayesian model averaging performed with the function `bic.glm` from the R package `BMA`. The first column of results represents the probability of inclusion, the second and the third the expected value and standard deviation. The remaining columns show the results of the most probable models. Variables which are significant (PIP higher than 0.5) are in bold letters.

variable and why it would be correlated to the dummy vairable and at the same time not be correctly proxied by GDP per capita.

The other main revelation that the BMA provides refers to Linder's variable. While this variable was not significant in the model including all the variables, its probability of inclusion is of 0.77 when considering all models, and it is included in the most likely model. Its expected value, around -0.01, is somewhat smaller in magnitude than the one that is found in Keum (2010) for aggregate trade, which is around -0.06 (although he obtained positive or non-significant effects when considering tourism).

The graphs of the estimated posterior densities of the coefficients are shown in figure 5.4 in the appendix. The black lines on zero show the cumulative probability of the models which do not include the variable, and therefore represents the probability of not inclusion $(1-PIP)$. The package also shows the conditional means and conditional standard deviation values as defined in equations 4.9 and 4.11.

In conclusion, the BMA is a useful procedure that sheds some light on the uncertainties that considering only one model can cause, while having a smoothing nature which should reduce concerns on specification bias, especially those arising from very high correlations. While it may not reveal new results regarding those determinants whose effect was already well assessed, it does help make stronger conclusions on the variables whose role was dubious and for which only one perspective might be non determinant.

Chapter 5

Conclusions

There are several conclusions that can be drawn from this work, referring both to the problem studied, the determinants of Italian domestic tourism, and to the choice and implementation of the PPML estimation and the BMA as a robustness test. Some aspects also need to be discussed in order to understand the results and to be able to contextualise them within the limits of this work.

First of all, it has been possible to corroborate how the classic tourism determinants have a significant effect on tourism flows. Distance has shown to be a stronger determinant than in the case of international tourism, which has been explained both from the point of view of costs (as international travels distance correlates less with monetary and time costs than in the case of national travels) and from the point of view of exotism, which likely plays a higher role for international flows. In fact, exotism is probably the main cause behind the positive coefficient of *island_dest*, showing that the difficulties associated to the travel to an island are more than compensated by the characteristic attractions of the islands compared to continental destinations. This effect does not function in the reverse direction, as the elasticity of *island_or* is negative. The problem of correctly dissociating the underlying competing causes which once aggregated lead to the value of the elasticity of distance has been very scarcely treated in the literature, and the matter of the differences between national and international flows is only a special case of this wider problem. Since distance is the key variable of the gravity model, and the most constant one in terms of how it is defined across papers, it is important that studies do not limit themselves to yield an estimate of the elasticity of flows to distance, but also try to compare it with other similar works and provide a reasoning for their similarities or discrepancies. Hopefully, this work has shed some light on the matter and has provided some reasoning to the results that have been found.

The other main variables proposed by the economic derivation, income and population at the origin and their destination counterparts, have also been shown to be significant. The higher elasticity of income at destination than at origin can be surprising. Still, I suspect that income at origin may have a complex effect on national tourism, as it probably greatly increases the absolute number of tourism flows from that region, but high incomes may cor-

relate with more international travels, therefore pulling downwards the coefficient. Perhaps, had I had access to data on international trips I would have obtained a different coefficient. Besides, there are a very high number of characteristics that can correlate with income at destination, of which only two have been studied. A future line of work on this matter should take use data of international travels and include more destination characteristics to account for these matters. Regarding population, the elasticity of one is exactly what was expected, suggesting that everything equal an increase of one percent of the population implies an increase of one percent of travel to every destination, while the smaller but significant elasticity of population at destination does probably reflect the influence of absolute number of attractions and facilities at the destination not already accounted for *culture*. Seeing the very high differences in these effects, I would strongly discourage the old practice (still sporadically followed today) of summing origin and destination populations yielding to a total population mass variable, since aggregating under such different elasticities can lead to strong bias and the impossibility of correctly interpreting results.

The culture and crime variables have had very consistent effects with the same sign as it was expected, which can be seen as a success, both in terms of the adequacy of the model and as a good sign regarding its implications for public policy interventions to increase the touristic appeal of a region. However, there remains the question of how much these coefficients reflect causality and how much they reflect correlation. I believe that it is safe to trust that *culture*'s coefficient reflects the actual attraction by historical sites and museums in a region, although there are probably differences in the magnitude of these effects comparing UNESCO heritage sites with more modest cultural institutions. Since the number of the latter is more easily increased than that of the first, this estimate should be understood as an almost inaccessible upper bound to what a local government can attain regarding culture expenditure. Again, it would be interesting in future works to disaggregate cultural attractions and compare the new etimates. On the other hand, it is possible that the crime rate correlates with other characteristics of a region such as inefficient public transportation, degradation of the public spaces, etc. Still, the inclusion of GDP per capita at the destination level should reduce the amount of bias arising from hidden variables and there is no reason to believe that crime rates do have a strong repulsion effect to tourism.

Another success of this work has been the confirmation of the effect of Linder's hypothesis also at the national level and for the case of tourism. Still, I would argue that the variable of economic distance probably shows the effect of more phenomena than that described by Linder's reasoning of similar factors' prices leading to similar services, and therefore attracting tourists from equally rich regions. In fact, an important part of Italian travels are motivated by labour (almost 8 million in 2019 according to ISTAT: *Viaggi per motivo* data), and it is likely that these flows are more common among regions with similar industries. Another study could focus on the differences of leisure and business travel, although flows disaggregated by motive are not available for Italy yet. Furthermore, it could even be argued that places with similar geographical characteristics have similar comparative advantages, therefore centering on similarly value-added industries. If tourists are attracted by the geographical and not the economic similarities, then Linder's variable would be proxying

more phenomena than that described by Linder's hypothesis. A similar reasoning could be made regarding the role of historical ties and cultural resemblances, which may both attract tourists and correlate with economic distance. In order to discern these possibly entangled effects, I would propose a bigger model where some variables could account for cultural and geographical ties such as belonging to the same political entity before the Italian unification, having similar elevations, temperatures, etc.

There remain the variables which have a less clear effect on tourism and which also need some more discussion. I have already mentioned that the *border* variable can have an effect contrary to what was expected due to a higher prevalence of one-day-trips in flows between adjacent regions. In order to be able to affirm this it would be advisable to study the model changing the dependent variable from stays to total number of leisure trips with or without staying at the destination. I would expect a strong change in the *border* coefficient in that case. Still, it is also possible that this contradictory result is caused by a failure of the specification to correctly account for the role of distance. Some recent works using the gravity model have allowed for more complex effects of distance on flows (for example Congdon (2000) where the author permits a stronger effect of physical distance for short distances and a growing effect of car distance for bigger distances), and it can be the case that its effect is not linear, but it plays a smaller effect for short distance and a bigger effect for large ones. Therefore the *border* variable effect would only be a correction of this bias. This theory seems to be backed from the decrease of the distance effect in model 4 of the table 4.1, which is the only of the most likely models not accounting for adjacency, compared to the rest. Fitting a model where distance has a polynomial and not only a linear effect could be informative on this regard.

Finally, the biggest surprise has come from the positive effect of price at the destination, which has been found when performing the PPML regression to the model and under the BMA results. The information obtained from running the alternative specifications was somewhat ambiguous, relative prices still have a positive effect when the MR variable is included and it becomes non significant when it is excluded. It seems that rather than an economic phenomenon that can be explained, it is the result of a misspecification arising from the inclusion of the MR term, although it is still possible, as already mentioned, that it can be the result of hidden variables bias. Due to this incongruity of prices, I would not venture to interpret the effect of MR even though the sign was the expected one. Instead, I would propose to take into account the multilateral resistance factor in a different way, ideally basing more strongly on the derivation from Anderson and Van Wincoop (2003) and making assumptions on the elasticity of substitution of tourists.

In general, it could be concluded that the PPML has been successful in estimating coherent elasticities for the different determinants, and the cases of discrepancy correspond more to problem arising from the selected variables than to the estimation procedure. In fact, the differences between the OLS and the PPML should be accepted as the result of the bias arising from the transformation of the dependent variable in the first case. Hopefully, the work has also helped introducing the reader to quasi-maximum likelihood estimation and has provided some guidance on its use. There are still many discrepancies and contradicting

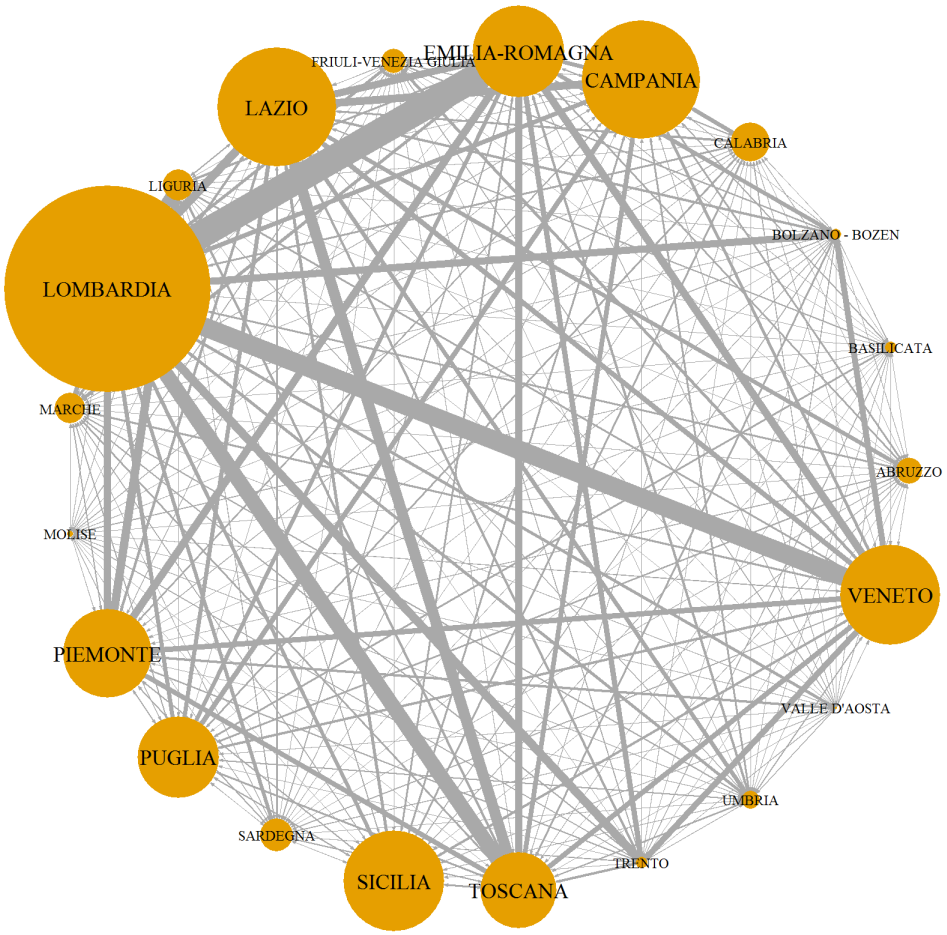
points of view regarding what is permitted and what it is not under its use, and some more theoretical work, and especially empirical guides, are necessary.

Regarding the model, there remains the possibility that not having included fixed effects has led to bias in the results. Indeed, some of the possible hidden variables that I have just identified could have theoretically been prevented by taking at least region-pairs fixed effects, and some even having taken only origin and destination. The decision to only include time effects has been justified by the will of included as many variables as possible to have a more comprehensive visual than otherwise it would have been possible, and it has also helped regarding the computational viability of the demanding technique that is the BMA. In fact, origin and destination time-varying fixed effects could have helped eliminate the multilateral resistance bias, but its implementation would have been so demanding and the set of possible variables included so restricted that the study would have lost its essence. Yet, for studies focusing on a reduced set of time-varying variables, I would follow the literature and suggest using a model with fixed effects

The last chapter preceding the conclusions has been an application of a very simple BMA procedure which has helped better identify the effects of the different determinants. Hopefully, the practice of studying more than one specification can become more common in the future. Even if a specification is clearly and uniquely defended by a theoretical derivation, the BMA keeps being a very informative robustness test that allows for verifying the significant coefficients and obtaining more information about the least clear ones. Furthermore, it also helps to understand how removing a set of variables influences the results obtained by the remaining ones, as it was the case with the *border* variable and *distance*. And instead of blindly perfrming this way, the BMA already identifies the most likely models, so that comparisons can be better focused. As I have already mentioned, the relation among the sets of variables can also be studied under a BMA perspective in a more structured way like in Beck et al. (2017b). Thus, I conclude by saying that the bayesian model averaging is a technique that offers multiple benefits to the researcher and whose possibilities have been barely touched in this work.

Appendix

Figure 5.1: Travel flows network in 2019



Note: The width of the edge is proportional to the number of travels summing both directions of the flow and the size of the vertices is proportional to the population of the region. The image has been created using the package `igraph` for defining the underlying graph structure, and `ggplot2` for the image design.

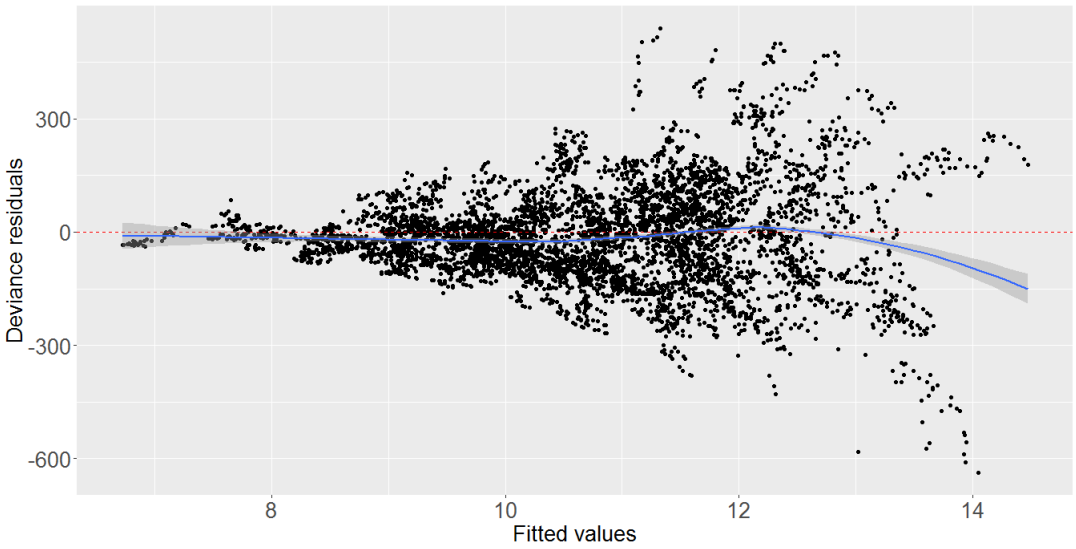
Table 5.1: Correlation matrix of all the explanatory variables.

	distance	GDP_pc_or	GDP_pc_dest	population_or	population_dest	linder	island_or	island_dest	border	crime	culture	CPI_or	CPI_dest	MR
distance	1.00	-0.19	-0.19	-0.02	-0.02	0.59	0.33	0.33	-0.49	-0.14	-0.07	0.08	0.08	-0.16
GDP_pc_or	-0.19	1.00	-0.03	0.05	-0.00	0.10	-0.37	0.02	0.07	-0.05	-0.01	-0.04	0.07	-0.14
GDP_pc_dest	-0.19	-0.03	1.00	-0.00	0.05	0.10	0.02	-0.37	0.08	0.26	0.14	0.07	-0.04	0.23
population_or	-0.02	0.05	-0.00	1.00	-0.05	-0.00	0.06	-0.00	0.08	-0.03	-0.03	-0.03	0.01	-0.01
population_dest	-0.02	-0.00	0.05	-0.05	1.00	-0.00	-0.00	0.06	0.08	0.61	0.68	0.01	-0.03	0.05
linder	0.59	0.10	0.10	-0.00	-0.00	1.00	0.03	0.03	-0.29	-0.11	-0.12	0.15	0.15	-0.19
island_or	0.33	-0.37	0.02	0.06	-0.00	0.03	1.00	-0.05	-0.13	0.01	-0.01	0.01	-0.00	-0.01
island_dest	0.33	0.02	-0.37	-0.00	0.06	0.03	-0.05	1.00	-0.13	-0.12	0.13	-0.00	0.01	-0.03
border	-0.49	0.07	0.08	0.08	0.08	-0.29	-0.13	-0.13	1.00	0.08	0.06	-0.03	-0.03	0.06
crime	-0.14	-0.05	0.26	-0.03	0.61	-0.11	0.01	-0.12	0.08	1.00	0.73	-0.12	-0.20	0.16
culture	-0.07	-0.01	0.14	-0.03	0.68	-0.12	-0.01	0.13	0.06	0.73	1.00	0.00	-0.02	0.03
CPI_or	0.08	-0.04	0.07	-0.03	0.01	0.15	0.01	-0.00	-0.03	-0.12	0.00	1.00	0.74	-0.06
CPI_dest	0.08	0.07	-0.04	0.01	-0.03	0.15	-0.00	0.01	-0.03	-0.20	-0.02	0.74	1.00	-0.13
MR	-0.16	-0.14	0.23	-0.01	0.05	-0.19	-0.01	-0.03	0.06	0.16	0.03	-0.06	-0.03	1.00

Note: Computed with the deafault R function `cor`.

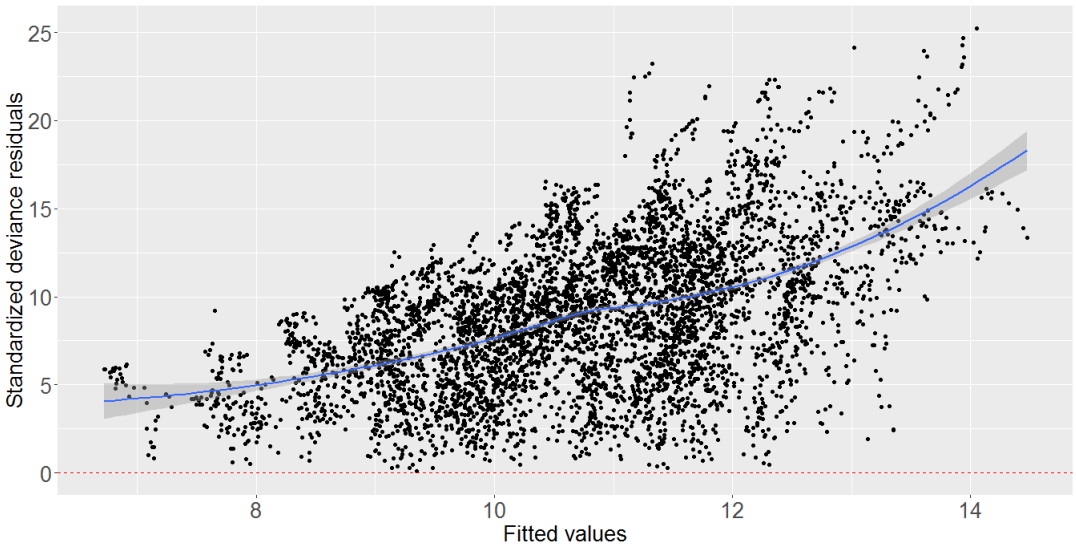
Figure 5.2: Residuals plots

Deviance residuals vs. fitted values



Note: The points correspond to the deviance residual of every observation for the PPML model with time effects. The plot has been created using commands from library `ggplot2` in R and obtaining the deviance residuals from the `residuals` function.

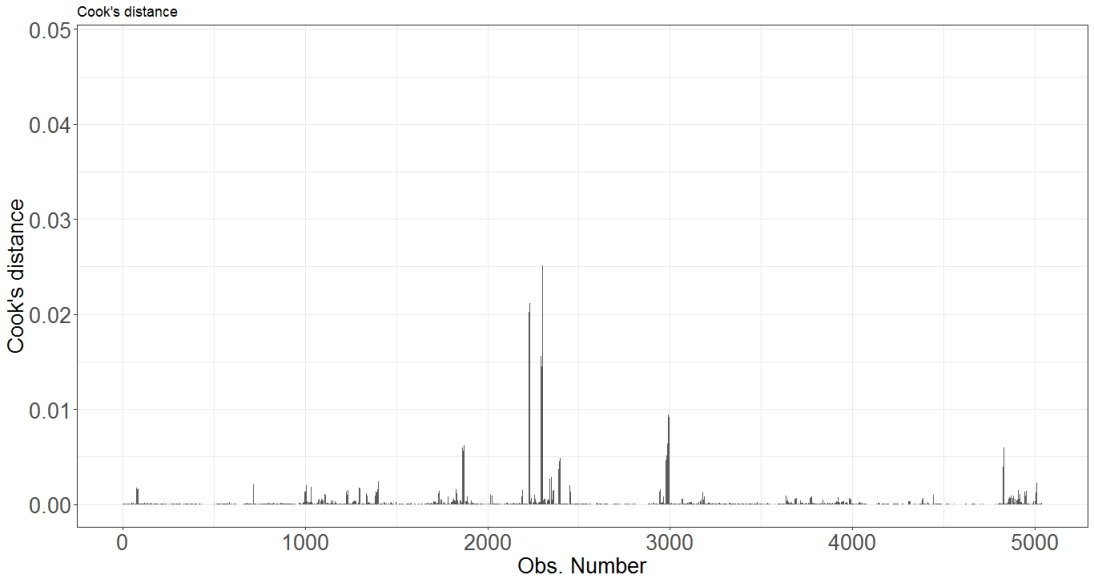
Standardised deviance residuals vs. fitted values



Note: The points correspond to root of the absolute value of the deviance residual of every observation for the PPML model with time effects. The plot has been created using commands from library `ggplot2` in R and obtaining the deviance residuals from the `residuals` function.

Figure 5.3: Outliers identification

Cook's distance plot



Residuals vs. leverage

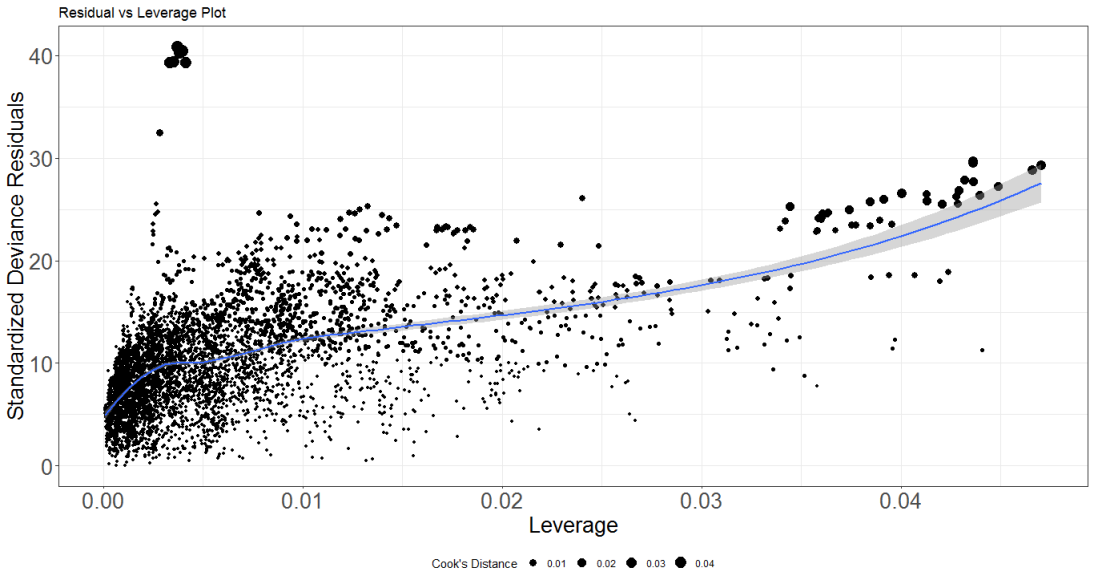


Table 5.2: Regression results of the model where CPI enters as rate of the destination over the origin.

Estimator: Dependent variable:	OLS $\ln X_{ijt} + 1$	OLS (Time eff.) $\ln X_{ijt} + 1$	PPML X_{ijt}	PPML (Time eff.) X_{ijt}
distance	-0.76*** (0.02)	-0.76*** (0.02)	-0.51*** (0.01)	-0.51*** (0.01)
GDP_pc_or	0.09* (0.04)	0.08* (0.04)	0.36*** (0.03)	0.35*** (0.03)
GDP_pc_dest 0.98***	0.97*** (0.04)	0.79 *** (0.04)	0.76 *** (0.03)	 (0.03)
population_or	1.03*** (0.01)	1.03*** (0.01)	1.04*** (0.01)	01.04*** (0.01)
population_dest	0.47*** (0.01)	0.46*** (0.01)	0.32*** (0.01)	0.31*** (0.01)
island_or	-0.43*** (0.03)	-0.43*** (0.03)	-0.27** (0.03)	-0.28*** (0.03)
island_dest	0.47*** (0.01)	0.46*** (0.01)	0.32*** (0.01)	0.31*** (0.01)
eu_border	0.05** (0.02)	0.06** (0.02)	0.03* (0.01)	0.04** (0.01)
border	-0.09*** (0.03)	-0.09*** (0.03)	-0.06*** (0.02)	-0.05** (0.02)
crime	-4.98*** (0.41)	-4.73*** (0.48)	-3.10*** (0.32)	-2-36*** (0.38)
culture	2.12*** (0.11)	2.11*** (0.11)	1.04** (0.09)	0.99*** (0.10)
linder	0.07*** (0.01)	0.07*** (0.01)	-0.01 (0.01)	-0.01 (0.01)
CPI_rate	5.09*** (0.24)	5.12*** (0.24)	0.98*** (0.23)	0.98*** (0.23)
MR	3.95*** (0.12)	3.96*** (0.12)	3.19*** (0.04)	3.20*** (0.04)
Year effects	No	Yes	No	Yes
Null deviance: Residual deviance: R^2 and pseudo R^2	 0.896	 0.896	846039967 84883390 0.900	846039967 84366618 0.900

Note: The OLS estimation has been performed using the `lm` command in R, while the PPML estimator has been applied using the `ppml` function from the package `Gravity` also in R. The estimated coefficients are shown in the same level of the variable name, while each estimated standard deviation is shown under it. The deviance and R^2 values are obtained applying the function `summary` to the respective models and the pseudo-R^2 was computed using the deviances. Stars denote p-values as follows: $^*p < 0.05$; $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 5.3: Regression results of the model with relative CPI and excluding the MR term.

Estimator: Dependent variable:	OLS $\ln X_{ijt} + 1$	OLS (Time eff.) $\ln X_{ijt} + 1$	PPML X_{ijt}	PPML (Time eff.) X_{ijt}
distance	-0.95*** (0.02)	-0.95*** (0.02)	-0.70*** (0.02)	-0.70*** (0.02)
GDP_pc_or	0.01 (0.04)	0.01 (0.04)	0.26*** (0.05)	0.26*** (0.05)
GDP_pc_dest 1.20***	1.20*** (0.04)	0.91*** (0.04)	0.90 *** (0.05)	 (0.03)
population_or	1.01*** (0.01)	1.01*** (0.01)	0.98*** (0.01)	0.98*** (0.01)
population_dest	0.58*** (0.01)	0.58*** (0.01)	0.43*** (0.01)	0.42*** (0.01)
island_or	-0.32*** (0.03)	-0.32*** (0.03)	-0.16** (0.03)	-0.16*** (0.03)
island_dest	0.30*** (0.03)	0.30*** (0.03)	0.09 (0.05)	0.09 (0.05)
eu_border	0.08*** (0.02)	0.08*** (0.02)	0.08*** (0.02)	0.08*** (0.02)
border	-0.02*** (0.03)	-0.02*** (0.03)	-0.04*** (0.03)	-0.04** (0.03)
crime	-4.81*** (0.45)	-4.94*** (0.53)	-3.54*** (0.51)	-3.24*** (0.60)
culture	2.01*** (0.12)	2.02*** (0.12)	1.15** (0.15)	1.13*** (0.15)
linder	0.08*** (0.01)	0.08*** (0.01)	-0.01 (0.01)	-0.01 (0.01)
CPI_rel	2.87*** (0.25)	2.86*** (0.26)	-0.14 (0.37)	-0.16 (0.37)
Year effects	No	Yes	No	Yes
Null deviance: Residual deviance: R^2 and pseudo R^2	 0.873	 138592650 0.873	846039967 138012691 0.836	846039967 0.837

Note: The OLS estimation has been performed using the `lm` command in R, while the PPML estimator has been applied using the `ppml` function from the package `Gravity` also in R. The estimated coefficients are shown in the same level of the variable name, while each estimated standard deviation is shown under it. The deviance and R^2 values are obtained applying the function `summary` to the respective models and the pseudo-R^2 was computed using the deviances. Stars denote p-values as follows: $^*p < 0.05$; $^{**}p < 0.01$, $^{***}p < 0.001$.

Figure 5.4: Plotted posterior densities of the elasticities coefficients.

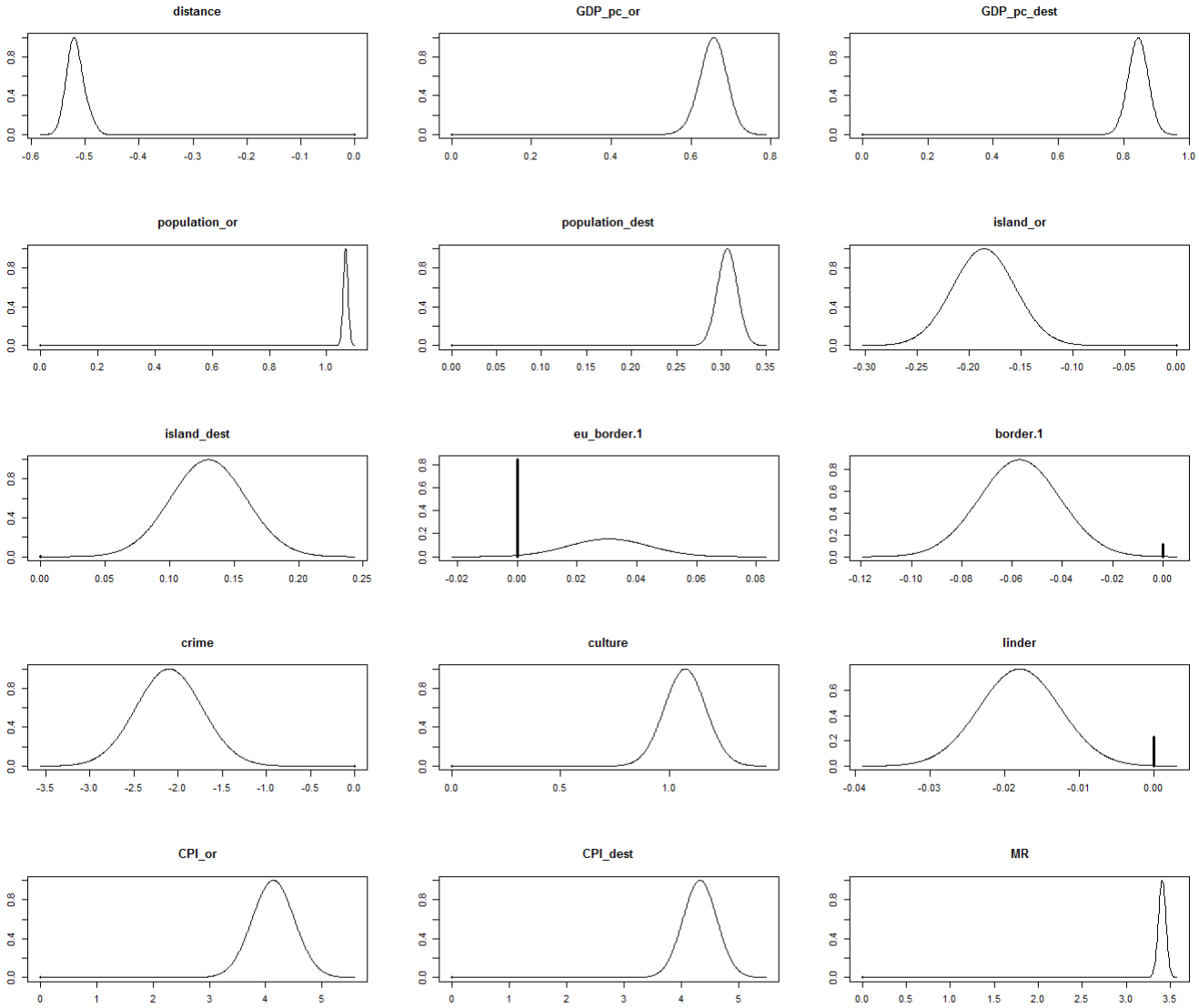


Table 5.4: Conditional expected value and standard deviation

	PIP	EV	EV cond.	SD	SD cond.
dist_log	100.00	-0.52	-0.52	0.02	0.02
GDP_pc_or	100.00	0.66	0.66	0.03	0.03
GDP_pc_dest	100.00	0.84	0.84	0.03	0.03
population_or	100.00	1.07	1.07	0.01	0.01
population_dest	100.00	0.31	0.31	0.01	0.01
island_or	100.00	-0.19	-0.19	0.03	0.03
island_dest	99.50	0.13	0.13	0.03	0.03
eu_border	11.70	0.00	0.03	0.01	0.01
border	88.00	-0.05	-0.06	0.02	0.02
crime	100.00	-2.10	-2.10	0.37	0.37
culture	100.00	1.07	1.07	0.10	0.10
linder	80.40	-0.01	-0.02	0.01	0.01
CPI_or	100.00	4.14	4.14	0.37	0.37
CPI_dest	100.00	4.33	4.33	0.29	0.29
MR	100.00	3.41	3.41	0.04	0.04

Bibliography

Amini, S. M., & Parmeter, C. F. (2011). Bayesian model averaging in r. *Journal of Economic and Social Measurement*, *36*(4), 253–287.

Anderson, J. E. (1979). A theoretical foundation for the gravity equation. *The American economic review*, *69*(1), 106–116.

Anderson, J. E. (2011). The gravity model. *Annu. Rev. Econ.*, *3*(1), 133–160.

Anderson, J. E., & Van Wincoop, E. (2003). Gravity with gravitas: A solution to the border puzzle. *American economic review*, *93*(1), 170–192.

Beck, K. et al. (2017a). Bayesian model averaging and jointness measures: Theoretical framework and application to the gravity model of trade. *Statistics in Transition. New Series*, *18*(3), 393–412.

Beck, K. et al. (2017b). Bayesian model averaging and jointness measures: Theoretical framework and application to the gravity model of trade. *Statistics in Transition. New Series*, *18*(3), 393–412.

Cameron, A. C., & Trivedi, P. K. (1990). Regression-based tests for overdispersion in the poisson model. *Journal of econometrics*, *46*(3), 347–364.

Chasapopoulos, P., Den Butter, F. A., & Mihaylov, E. (2014). Demand for tourism in greece: A panel data analysis using the gravity model. *International Journal of Tourism Policy*, *5*(3), 173–191.

Chen, H., Mirestean, A., & Tsangarides, C. G. (2018). Bayesian model averaging for dynamic panels with an application to a trade gravity model. *Econometric Reviews*, *37*(7), 777–805.

Cochrane, R. (1975). A possible economic basis for the gravity model. *Journal of Transport Economics and Policy*, 34–49.

Congdon, P. (2000). A bayesian approach to prediction using the gravity model, with an application to patient flow modeling. *Geographical analysis*, *32*(3), 205–224.

Culiuc, M. A. (2014). *Determinants of international tourism*. International Monetary Fund.

De Benedictis, L., & Taglioni, D. (2011). The gravity model in international trade. *The trade impact of european union preferential policies* (pp. 55–89). Springer.

Dunn, P. K., & Smyth, G. K. (2018). *Generalized linear models with examples in r*. Springer.

Durbarry, R. (2008). Tourism taxes: Implications for tourism demand in the uk. *Review of Development Economics*, *12*(1), 21–36.

Heinzl, H., & Mittlböck, M. (2003). Pseudo r-squared measures for poisson regression models with over-or underdispersion. *Computational statistics & data analysis*, *44*(1-2), 253–271.

Isard, W. (1954). Location theory and trade theory: Short-run analysis. *The Quarterly Journal of Economics*, *68*(2), 305–320. http://www.jstor.org/stable/1884452

Keum, K. (2010). Tourism flows and trade theory: A panel data analysis with the gravity model. *The Annals of Regional Science*, *44*(3), 541–557.

Khadaroo, J., & Seetanah, B. (2008). The role of transport infrastructure in international tourism development: A gravity model approach. *Tourism management*, *29*(5), 831–840.

Kimura, F., & Lee, H.-H. (2006). The gravity equation in international trade in services. *Review of world economics*, *142*(1), 92–121.

Linder, S. B. (1961). *An essay on trade and transformation*. Almqvist & Wiksell Stockholm.

Liu, E. (2019). *Deviance of poisson regression. master. australia.*

Massidda, C., & Etzo, I. (2012). The determinants of italian domestic tourism: A panel data analysis. *Tourism Management*, *33*(3), 603–610.

Mayo, E. J., Jarvis, L. P., & Xander, J. A. (1988). Beyond the gravity model. *Journal of the Academy of Marketing Science*, *16*(3), 23–29.

Moral-Benito, E. (2011). Model averaging in economics.

Morley, C., Rosselló, J., & Santana-Gallego, M. (2014). Gravity models for tourism demand: Theory and use. *Annals of Tourism Research*, *48*, 1–10.

Petrella, A., Torrini, R., Barone, G., Beretta, E., Breda, E., Cappariello, R., Ciaccio, G., Conti, L., David, F., Degasperi, P., et al. (2019). *Turismo in italia: Numeri e potenziale di sviluppo*. Banca d'Italia.

Priego, F. J., Rosselló, J., & Santana-Gallego, M. (2015). The impact of climate change on domestic tourism: A gravity model for spain. *Regional environmental change*, *15*(2), 291–300.

Raftery, A., Painter, I., & Volinsky, C. (2005). Bma: An r package for bayesian model averaging. *R News*.

Ranjan, P., & Tobias, J. L. (2007). Bayesian inference for the gravity model. *Journal of Applied Econometrics*, *22*(4), 817–838.

Shahriar, S., Qian, L., Kea, S., & Abdullahi, N. M. (2019). The gravity model of trade: A theoretical perspective. *Review of Innovation and Competitiveness: A Journal of Economic and Social Research*, *5*(1), 21–42.

Silva, J. S., & Tenreyro, S. (2006). The log of gravity. *The Review of Economics and statistics*, *88*(4), 641–658.

Stewart, J. Q. (1948). Demographic gravitation: Evidence and applications. *Sociometry*, *11*(1/2), 31–58.

Tinbergen, J. (1962). Shaping the world economy; suggestions for an international economic policy.

Tyazhelnikov, V., & Zhou, X. (2020). Ppml, gravity, and heterogeneous trade elasticities.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the econometric society*, 1–25.