STUDIECENTRUM VOOR ECONOMISCH EN SOCIAAL ONDERZOEK

INTERPRETING COEFFICIENTS IN A GRAVITY MODEL

Dr. G. BLAUWENS

werknota 7532/10291

januari 1975

Universitaire Faculteiten St.-Ignatius
Prinsstraat 13 - 2000 Antwerpen
D/1975/1169/01

SESO
STUDIECENTRUM VOOR ECONOMISCH
EN SOCIALE ONDERZOEK
SINT-IGNATIUS, PRINSSTRAAT 13, ANTWERPEN

INTERPRETING COEFFICIENTS IN A GRAVITY MODEL

Dr. G. Blauwens

Since their development in the early fourties (°), the popularity of gravity models has been continuously growing. Though originally conceived to explain passenger movements, these models have also been used in other fields, such as commodity transportation (°°) and location theory (°°°). Gravity models, also called interactance or potential models, are at present a standard tool for transportation economists.

The development of the underlying theory however has not kept pace with practical applications. While remarkable statistical fit is being reported regularly, the theory remains vague and obscure. The parameters in the model are estimated with high statistical accuracy but their economic meaning is not clear. To some extent the statistical success of gravity models is even due to the vagueness of the theory, allowing almost unlimited experiments with the specification.

(°) S. STOUFFER, "Intervening opportunities: A theory relating mobility and distance", American Sociological Review , Vol.5, nr.4, Dec. 1940, pp. 845-867.

G.K. ZIPF, "The P_1P_2/D Hypothesis: The case of Railway Express", Journal of Psychology, 11, 1946.

J.Q. STEWART, "Empirical Mathematical Rules Concerning the Distribution and Equilibrium of Population", Geographical Review, July 1947, pp. 461-485.

(°°) E.g. J.R. MEYER & M.R. STRASZHEIM, Techniques of Transport Planning, Vol.1: Pricing and Project Evaluation, Washington, 1971, pp. 165-169. (Further references in this book on pp.326-328), T. PEETERS, "Application du modèle gravitationnel à la structure des échanges internationaux de biens d'équipement", Statistische Studies en Enquêtes, 1970, nr.1, pp. 63-94, and last but not least H. LINNEMAN, An econometric study of international trade flows, Amsterdam, 1966.

(°°°) E.g. W. KAU, Theorie und Anwendung raumwirtschaftlicher Potentialmodelle, Tübingen, 1970.

A considerable number of authors (°) have stressed this theoretical weakness.
We should admit that the development of gravity models has been inspired by doubtful assumptions on the probability of interaction and by dubious analogies to fysical gravitation laws (°°).

However, when criticising a model, its purpose should be kept in mind. Objections which are correct when a model is used as a forecasting device, are not necessarily correct when the aim of the model is parameter estimation. Further, a distinction has to be made between different specifications. If one variant of the gravity model presents serious drawbacks, another variant may still be perfect.

In this paper we will discuss gravity models as a tool for estimating the price elasticity of transport demand. In §1 we sketch a spatial equilibrium framework. In §2 the specification of gravity models is briefly restated. Next, in §3, we try to establish a link between the economic equilibrium system of §1 and the highly simplified specification of gravity models. This will allow an economic interpretation of the elasticities, estimated by these models. Paragraph 4 warns for some traps, which should be avoided when specifying and interpreting gravity models. Pragraph 5 finally contains an application to Belgian transport flows.

(°) Typical examples are G.K. SLETMO, Demand for air cargo, Bergen 1972, p.21, J.M. HARTWICK, "The gravity hypothesis and transportation cost minimization", Regional and Urban Economics, 1972, pp.297-308, I.G. HEGGIE, "Are Gravity and Interactance Models a Valid Technique for Planning Regional Transport Facilities?", Operational Research Quarterly, Vol.20, nr.1, pp.93-110, and I.S. JONES, "Gravity Models and Generated Traffic", Journal of Transport Economics and Policy, Vol.4, nr.2, 1970, pp.208-211.

(°°) Cfr. G.A.P. CARROTHERS, "An historical review of the gravity and potential concepts of human interaction", Journal of the American Institute of Planners, 1956, vol.22, pp.94-102, W. ISARD, Methods of Regional Analysis, New York,-London, 1960, pp.494-504. Fore more recent developments, cfr. A.G. WILSON, "The use of entropy maximising methods", Journal of Transport Economics and Policy, vol.III, 1969, pp.108-112.

§1. SPATIAL EQUILIBRIUM FLOWS

When we divide the economic space into I regions, we can distinguish I^2 transport flows from a region of origin to a region of destination. Let us represent these flows with the symbol x_{ij} (i,j=1...I). The index i indicates the origin and j represents the destination. The ordered pair ij denotes the "relation".

Depending on the case which is being analysed, x_{ij} can represent the daily number of passengers travelling from i to j, it can mean the number of commuters from i to j during the morning peak, it can be the yearly tonnage of chemical products transported from i to j, it can be the monthly total of truckloads from i to j, etc.

An equilibrium mechanism, defining the I^2 variables x_{ij}, is given by the following set of equations:

1.a. Demand equations

The demand for transportation depends, among other things, on the monetary cost and on the time cost of transportation. This dependence is quite complex, as a flow x_{ij} is not only influenced by cost and time on relation ij itself, but also by cost and time on other relations. As a matter of fact, the desirability of transporting from i to j is highly dependent on the cost of alternative relations. The competitiveness of region i as a supplier of region j clearly depends on the costs of sypplying j from alternative origins. Tourist travel or even commuting to a destination j may clearly be affected when another destination comes within cheaper reach.

When we indicate the monetary cost of transportation from i to j with the symbol c_{ij} and the matrix of I^2 elements c_{ij} (i,j=1...I) with C, we can say that in general a single transport flow x_{ij} will depend on all the elements of the matrix C, i.e. on I^2 transport costs c_{ij}. It will also depend on I^2 transport times t_{ij}, i.e.

on the I^2 elements of a matrix T. We can represent this concisely with the expression

$$x_{ij} = f_{ij}(T,C) \qquad (i,j=1...I)$$

meaning that the demand for transport on a given relation ij is a function f_{ij} of the $2I^2$ elements contained in the time matrix T and the cost matrix C.

These $2I^2$-variable functions are still less complex than reality. A cost c_{ij} or a transit time t_{ij} is not well defined in the real world. One single relation can present different transport tariffs and times according to the mode which is being used or to the different transport categories which we have summed in the aggregates x_{ij}.

In general, we can state that each relation ij presents K different transport costs c_{ij}^1, c_{ij}^2, ..., c_{ij}^K and K different transport times t_{ij}^1, t_{ij}^2, ..., t_{ij}^K, all refering to a specific mode and a specific transport category. Introducing this in the demand functions, we obtain

$$x_{ij} = f_{ij}(T^1,T^2,...,T^K,C^1,C^2,...,C^K) \qquad (i,j=1...I) \quad (1)$$

meaning that transport demand on a given relation ij is a $2KI^2$-variable function of all the elements in the K time-matrices and the K cost-matrices.

Note that the I^2 demand functions (1) are all different functions, as expressed by adding the subscripts ij to the symbol f. Indeed, when we fix all the transport costs and times in the whole economy t_{ij}^k, c_{ij}^k (i,j=1...I, k=1...K) we will certainly not observe the same transport demand x_{ij} on all the relations. This is very obvious. The effect of a transport cost c_{ij}^k upon its "own" transport flow x_{ij} differs from its effect upon quite another flow x_{uv} (uv≠ij), if not in sign, certainly in magnitude!

It can also be noted at this stage that the demand functions (1) might in principle be estimated by means of time-series analysis. The consecutive observations of the flows x_{ij} on a given relation should in that case be regressed upon the $2KI^2$ explanatory variables in (1) plus exogeneous variables, accounting for shifts of f_{ij} over time. Of course, this would require a huge number of observations (at least $2KI^2$ plus a remainder to provide some degrees of freedom). Even if the specification in (1) is drastically simplified, time series of suitable length will rarely be available. In many applications the flows x_{ij} or the explanatory variables are known at only one moment in time. In these cases estimation by means of time series is excluded. Gravity models however may still work, as we will show later.

1.b. Cost equations (supply equations)

The monetary cost c_{ij}^k of transporting a given category of goods or passengers by a given mode from i to j, is not independent from the transport flows themselves. We should express this in cost equations of the following type

$$c_{ij}^k = g_{ij}^k (X) \qquad (i,j=1...I, k=1...K) \qquad (2)$$

where X is the IxI matrix of transport flows x_{ij} and where g_{ij}^k is an I^2-variable function, giving the transport cost as dependent upon the I^2 different flows in X. Indeed, the cost of transporting from i to j, depends not only on the flow x_{ij} itself, but also on other flows x_{uv} $(uv \neq ij)$, occuring elsewhere in space but using capacity of the same transport industry.

On the analogy of the demand functions, it is obvious that the KI^2 cost functions (2) are all different functions, as expressed by adding the indices i,j and k. Some relations have poor transport facilities and are sooner congested than others. Some relations present a long distance and consequently are more expensive than relations with a short distance.

1.c. Time equations

The equations giving transport time as a function of the flows, are completely analogous to the cost equations

$$t_{ij}^k = h_{ij}^k(X) \qquad (i,j=1...I, k=1...K) \qquad (3)$$

1.d. The equilibrium

The equilibrium system of demand, cost and time equations

$$x_{ij} = f_{ij}(T^1, T^2, \ldots, T^K, C^1, C^2, \ldots, C^K) \quad (i,j=1...I) \qquad (1)$$
$$c_{ij}^k = g_{ij}^k(X) \qquad (i,j=1...I, k=1...K) \qquad (2)$$
$$t_{ij}^k = h_{ij}^k(X) \qquad (i,j=1...I, k=1...K) \qquad (3)$$

contains as many equations as unknowns. It is complete. We cannot pretend it to be general and to contain all the possible details an analyst could wish: the system is essentially static, all its variables being measured at the same time, and it does not contain exogeneous variables, the latter being implicitely treated as parameters in the functions f_{ij}, g_{ij}^k and h_{ij}^k.

One can show however that time-series analysis and interregional programming, which are sometimes proposed as better alternatives for the gravity model, are nothing more than highly simplified special cases of our equilibrium system (°). The generality and detail of our system are thus sufficient as a framework for discussing gravity models.

(°) In interregional programming, even in very sophisticated variants, the equilibrium system is simplified by assuming g_{ij}^k and h_{ij}^k ($i,j=1...I, k=1...K$) to be constants, by lumping together time and monetary costs for different modes and transport categories and defining f_{ij} by means of some programming (minimum cost, maximum benefit) solution. One can verify this in P.A. SAMUELSON, "Spatial Price Equilibrium and Linear Programming", American Economic Review, Vol. XLII, nr.3, pp.283-303; TAKAYAMA T. & JUDGE G.G., "Spatial Equilibrium and Quadratic Programming", Journal of Farm Economics, Vol.46, February 1964, pp.67-93, Id., "Equilibrium among spatially separated markets: A Reformulation", Econometrica, Vol.32, nr.4, October 1964,

.../...

§2. GRAVITY MODELS

A very simple specification of gravity models is:

$$x_{ij} = O_i . D_j . \beta_o . (c_{ij})^{\beta_1} . e^{u_{ij}} \quad (i,j=1...I) \qquad (4)$$

with O_i as "origin potential" of region i, expressing its importance as an origin region and D_j as "destination potential" of region j, expressing its importance as a destination region (when the flows x_{ij} consist of morning peak commuters, O_i is population of region i and D_j is employment in region j. In a study of total freight flows, O_i can be production of region i and D_j total imports of region j, etc. In practical applications a wide variety of potentials are used.);

with c_{ij} the transport cost from i to j,

with e the basis of natural logarithms and $e^{u_{ij}}$ a disturbance (introduced in multiplicative form);

with β_o and β_1 parameters to be estimated.

.../... pp.510-524; TAKAYAMA T. & WOODLAND A.D., "Equivalence of Price and Quantity Formulations of Spatial Equilibrium", Econometrica, Vol.38, nr.6, November 1970, pp.889-906; Leon N. MOSES, "A general equilibrium model of production, interregional trade and location of industry", Review of Economics and Statistics, 1960, November, pp. 373-397; V.L. SMITH, "Minimization of Economic Rent in Spatial Price Equilibrium", Review of Economic Studies, Vol.XXX, nr.1, 1963, pp. 24-31; E. SILBERBERG, The demand for waterway transportation, Ann Arbor, 1964, pp.44-47; A. GHOSH & A. CHAKRAVARTI, "The problem of location of an industrial complex" in A.P. CARTER & A. BRODY, ed., Contributions to input-output analysis, Amsterdam-London, 1970, pp. 164-179; A.P. HURTER & L.N. MOSES, "Regional Investment and Interregional Programming", Papers of the Regional Science Association, Vol.13, 1964, pp.105-119; R. HOWES "A test of a linear programming model of agriculture", Papers of the Regional Science Association, Vol.19, 1967, pp.123-140; E. HEADY & A. EGBERT, "Programming Models of Interdependence among agricultural sectors and spatial allocation of crop production", Journal of Regional Science, 1962, Vol.2, nr.4, pp.1-20; M. L. GREENHUT, "Interregional Programming and the demand factor of location", Journal of Regional Science, 1967.

After a logarithmic transformation,

$$\ln \frac{x_{ij}}{O_i \cdot D_j} = \ln\beta_o + \beta_1 \ln c_{ij} + u_{ij}, \qquad (4)$$

the parameters $\ln\beta_o$ and β_1 can be estimated by means of standard linear regression techniques. The observations for this regression can in principle be provided by time series or by a cross-section or by a combination of both. In practice however the parameters are nearly always estimated from a cross-section of flows x_{ij}, occuring between different regions at the same time. In the subsequent discussion we will assume an estimation from such a cross-section and not from time series. This detail is important.

The difference between the very simple specification (4) and the rather detailed equilibrium system, described by equations (1), (2), (3), is enormous. This causes severe theoretical problems. Can the coefficient β_1 be interpreted as the price elasticity of transport demand? There are serious doubts about the validity of such an interpretation:

First, we face simultaneous equation problems. The observations on x_{ij} and c_{ij} can trace out not only the demand function (1) which we try to estimate, but also the cost function (2) or a mixture of both: the estimate of β_1 may contain both the effect of costs upon flows and the effect of flows upon costs.

Second, what exactly is meant by the transport cost c_{ij}? Is it an average of the different transport costs c_{ij}^1, c_{ij}^2, ..., c_{ij}^K occuring on this relation? Does it contain time costs? Is it simply a proxy such as geographical distance? In the latter case, does β_1 give the purely imaginary effect of increasing distance, that means tearing the earth apart?

Third, is there any sense in assuming the existence of one single function (4) and supposing it to be valid in a cross-section of

transport flows x_{ij} (i,j=1...I)? As we have seen, the demand functions (1) are different on different relations. In other words, there is no price elasticity β_1 for transport demand on all relations ij. Each relation reacts differently and possesses its own price elasticity.

Fourth, what exactly is contained in the disturbance term? Is it reasonable to assume zero covariance between this disturbance and the explanatory variable c_{ij}?

Fifth, why do we use potentials? What potentials should be used? The list of questions can be extended at pleasure. A variant of gravity models which is more sophisticated than the very simple example (4), is the following well-known specification:

$$x_{ij} = O_i \frac{D_j (c_{ij})^{\beta_1}}{\sum_{v=1}^{I} D_v (c_{iv})^{\beta_1}} \cdot e^{u_{ij}} \qquad (i,j=1...I) \quad (5)$$

making x_{ij} dependent not only upon its own transport cost c_{ij} but also on costs c_{iv} ($v \neq j$) of transporting from i to other regions. This is a step closer to the general demand functions (1).

Further refinement is sometimes introduced by distinguishing several transport costs, instead of one single cost figure c_{ij} per relation. This can lead to specifications of the types (6) and (7).

$$x_{ij} = O_i . D_j . \beta_o (c_{ij}^1)^{\beta_1} . (c_{ij}^2)^{\beta_2} \ldots (c_{ij}^K)^{\beta_K} . e^{u_{ij}} \qquad (i,j=1...I) \quad (6)$$

$$x_{ij} = O_i . \frac{D_j (c_{ij}^1)^{\beta_1} . (c_{ij}^2)^{\beta_2} \ldots (c_{ij}^K)^{\beta_K}}{\sum_{v=1}^{I} D_v . (c_{iv}^1)^{\beta_1} . (c_{iv}^2)^{\beta_2} \ldots (c_{iv}^K)^{\beta_K}} e^{u_{ij}} \qquad (i,j=1...I) \quad (7)$$

Refinements of this kind are once again a step closer to the general demand functions (1). Interpretation of the parameters however remains difficult. There is still a great difference between specifications like (6) and (7) on one hand, and the general demand functions (1) or the whole equilibrium system (1), (2), (3) on the other.

§3. DERIVATION OF GRAVITY MODELS

In this paragraph we will show how a gravity model can be obtained by introducing suitable assumptions or simplifications into the general demand functions (1). We will extensively treat the case of the simplistic specification (4). This is the most difficult example to handle, as it contains the highest degree of simplification. Once the derivation of (4) has been illustrated, the extention to other, less simplified, gravity models is straightforward. The steps for obtaining these specifications will be summarized briefly at the end of the paragraph.

Let us restate the general demand functions (1):

$$x_{ij} = f_{ij}(T^1, T^2, \ldots, T^K, C^1, C^2, \ldots, C^K) \quad (i,j=1\ldots I) \qquad (1)$$

These highly detailed functions explain the transport demand on a given relation by all the transport costs and -times, occuring in the whole economy. Of course, no econometrician can afford to handle equations of the complexity. No econometric model is able to explain a variable by all the variables which in reality affect it. Some variables are always treated as explanatory, some are always neglected and treated as a disturbance. The selection of explanatory variables should obey some rules:

1) It should contain these variables which are most interesting. (In prediction, the variables which have the most significant effect upon the dependent variable. In parameter estimation, at least all the variables which are associated with the parameters to be estimated.)

2) It should contain these variables which, when omitted, would cause correlation between the explanatory variables and the disturbance.

When estimating elasticities of transport demand, our main concern might be the direct price effect on a relation itself, while the estimation of cross price effects (effects of c^k_{uv} or t^k_{uv} upon x_{ij} ($uv \neq ij$)) might not be our purpose. In that case we can select one single element from each of the 2K matrices in (1) as explanatory and treat the other elements as a disturbance. This would give:

$$x_{ij} = \phi_{ij}(t^1_{ij}, t^2_{ij}, \ldots, t^K_{ij}, c^1_{ij}, c^2_{ij}, \ldots, c^K_{ij}) + \text{disturbance} \qquad (i,j=1\ldots I) \qquad (8)$$

The number of explanatory variables can still be reduced. For instance, some variables could be considered negligible and transferred to the disturbance. Or the explanatory variables could be substituted by a common proxy or by a generalised transport cost c_{ij}, the weights in this generalised transport cost being proportional to the coefficients of its components in ϕ_{ij}, etc. In the simplest case, this produces:

$$x_{ij} = \Phi_{ij}(c_{ij}) + \text{disturbance} \qquad (i,j=1\ldots I) \qquad (9).$$

As far, nothing unusual has happened. The acceptance of a disturbance term and the treatment of various prices by a common index, are standard procedures, used daily in econometrics. Admittedly, we do have to wonder whether the disturbance, containing the omitted cross-price effects, is uncorrelated with c_{ij}. At first sight however, it seems reasonable to accept this hypothesis. At closer examination there may appear some problems, but we will postpone this discussion. For the time being, we will accept the standard assumption of zero-covariance between c_{ij} and the disturbance term.

The main problem in deriving gravity models has not been solved yet: the function (9) still has the unpleasant property of being defined separately for each relation, as indicated by the subscripts i and j. Each relation has its own demand function Φ_{ij}, which is different from the demand function Φ_{uv} of another relation. On each relation, transport demand reacts differently upon transport cost. This rules out estimation of (9) from a cross-section of transport flows, observed between different regions at the same time. Such a cross-section contains exactly one observation for each of the functions (9). Each demand equation is observed in one point. It is out of the question, to estimate a function from this single observation.

To solve this problem, it would be easy to assume away the structural difference between relations, and to write (9) in the simplified form

$$x_{ij} = \Phi(c_{ij}) + \text{disturbance}$$

which in a cross-section of I^2 transport flows is observed I^2 times and can thus be estimated. Of course, the assumption of an identical demand function on all relations is highly unrealistic. No doubt, most of the opponents of gravity models have this radical assumption in mind. The derivation of gravity models however does not have to be put in the drastical terms, used above:

As a first step, which is not quite necessary in theory but will improve the statistical performance of the model, let us introduce potentials into the demand functions (9):

$$x_{ij} = O_i.D_j.F_{ij}(c_{ij}) + \text{disturbance} \qquad (i,j=1...I) \qquad (10)$$

where $F_{ij} = \Phi_{ij}/O_i.D_j$.

This does not change the specification. The only result is a reduction of the difference between demand functions on various relations. The functions F_{ij} (i,j=1...I) show less mutual difference than the original functions Φ_{ij}, as we have eliminated two factors, differing between regions and obviously explaining the expected value of x_{ij}.
We certainly do not pretend that all differences between demand functions have disappeared now and that all functions F_{ij} are identical. We only pretend that for estimation purposes the different F_{ij} can be replaced by a single function F and that the estimates of this function do give information on the individual F_{ij}.

To illustrate this, let us choose a functional form in (10), for instance, a loglinear one:

$$\ln\{\frac{x_{ij}}{O_i.D_j}\} = \ln\beta_o^{ij} + \beta_1^{ij} \ln c_{ij} + \text{disturbance} \quad (i,j=1...I) \qquad (11)$$

This is identical with the most simplified specification of gravity models (4), except for the fact that on each relation transport demand has a different constant and a different elasticity. Instead of the parameters β_o and β_1 in (4), we have in (11) β_o^{ij} and β_1^{ij}, differing between relations, as indicated by the superscripts.

One can show however that, apart from an "aggregation bias", the least squares estimate of $\ln\beta_o$ in (4) gives the mean of the various $\ln\beta_o^{ij}$, while with β_1 we estimate the mean of the various price elasticities β_1^{ij}.

A proof of this property can be given along the lines of general aggregation analysis, as developed by THEIL and KLOEK (°). A simpler proof runs as follows:

(°) H.THEIL, Linear Aggregation of Economic Relations, Amsterdam, 1954, Id., Principles of Econometrics, Amsterdam-London, 1971, pp.556-562. Also T.KLOEK, "Note on convenient matrix notations in multivariate statistical analysis", International Economic Review, Vol.2, 1961, pp.351-360. The proof in the case of gravity models can be found in G. BLAUWENS, De prijselasticiteit van de vraag naar goederenvervoer, (dissertation), Antwerp 1973, pp.75-77.

Define $\overline{\ln\beta_o}$ as the mean $\frac{1}{I^2}\sum_{i=1}^{I}\sum_{j=1}^{I}\ln\beta_o^{ij}$

$\overline{\beta_1}$ as the mean $\frac{1}{I^2}\sum_{i=1}^{I}\sum_{j=1}^{I}\beta_1^{ij}$

Add and substract these mean values in (11) to obtain:

$$\ln\{\frac{x_{ij}}{O_i.D_j}\} = \overline{\ln\beta_o} + \overline{\beta_1}\ln c_{ij} + \left[\text{disturbance} + (\ln\beta_o^{ij} - \overline{\ln\beta_o}) + (\beta_1^{ij} - \overline{\beta_1})\ln c_{ij}\right] \quad (i,j=1\ldots I) \quad (12)$$

The new disturbance, written between square brackets, contains the original disturbance of (11) plus additional terms, representing structural differences between relations. The inclusion of these additional terms is not so dramatic at all. The deviations of the parameters $\ln\beta_o^{ij}$ and β_1^{ij} from their means, are zero in expected value.

Hence, we only have to suppose their zero covariance with c_{ij}, to keep the disturbance zero in expected value and uncorrelated with the explanatory variables. Under this - reasonable - assumption the least-squares estimates of $\ln\beta_o$ and β_1 in the gravity model (4):

$$\ln\{\frac{x_{ij}}{O_i.D_j}\} = \ln\beta_o + \beta_1\ln c_{ij} + u_{ij} \qquad (i,j=1\ldots I) \quad (4)$$

are unbiased estimates of the mean values $\overline{\ln\beta_o}$ and $\overline{\beta_1}$.

The substitution of different constants β_o^{ij} by one single β_o and of different β_1^{ij} by one single elasticity β_1 seems a dramatic step at first sight. Upon closer examination, it is not so heroic at all: It leaves us with a specification that can be clearly interpreted, the constant being the mean of constants on the various relations, the elasticity being the mean of elasticities on the various relations. The estimates of these mean values can be considered as unbiased estimates, unless the elasticities are correlated with the transport cost c_{ij}. One has little reason however to suspect this, except when the functional form of the true demand functions is not constant elastic. In that case however, one should only change this functional form in (4).

The disturbance term in the simple specification (4) also has a clear economic interpretation: it contains the effect of neglected variables which are present in the general demand functions (1) but omitted in (4). It contains imperfections in the functional form and structural differences between the demand functions of the various relations. We know the content of the disturbance and can discuss its properties.

Finally, we understand the role of the potentials. Their presence in the spedification is not a theoretical necessity. Even when all the potentials O_i and D_j (i,j=1...I) are set equal to one and hence disappear from the specification, the different functions Φ_{ij} can be estimated by a single Φ without invalidating the derivations above. The potentials merely serve to eliminate as much structural differences from the disturbance term as possible. This results in smaller disturbance variance and hence in more accurate estimation.

We have shown how to obtain the specific variant (4) of the gravity model. Other variants can be obtained analogously in the following steps:

1) In the general demand functions (1), choose the appropriate explanatory variables and treat the other ones as a disturbance (compare (8) or (9));
2) Introduce suitable potentials, such as to minimise the remaining difference between demand functions on distinct relations (compare (10));
3) Choose a functional form (compare (11));
4) Estimate as if the demand functions were identical on all relations and obtain in this way their coefficients, averaged over all relations. (Compare (12)).

§4. SOME TRAPS TO BE AVOIDED

The coefficient estimates in a gravity model are unbiased estimates of the mean of corresponding coefficients in different demand equations. This has been shown in the preceding paragraph to be a correct statement, provided the disturbance term in the gravity model is not correlated with the explanatory variables. The latter assumption is highly important and will be examined now. Four points deserve further comments: simultaneous equations, intermodal competition, interspatial competition and aggregation bias.

4.1. Simultaneous equations

Correlation between the disturbance term and the explanatory variables in the gravity model, may be caused by simultaneous equations. Beside the demand functions (1), giving the effect of costs upon demand, the economic equilibrium also contains cost functions (2) and (3) (supply functions), giving the effect of the flows upon costs and times. Transport costs and times are not predetermined, but under the influence of (2) and (3), are correlated with the disturbance in the demand functions.

Of course, this simultaneous equations problem ceases to exist when geographical distance is used as a proxy for transport cost. This variable is clearly exogeneous. The problem only emerges when real transport prices and times are used.

An elegant solution is provided by the fact that the model is estimated from a cross-section of different relations, characterised by different geographical distance. This distance variable does not occur in the demand functions (1): Costs and times having been introduced already, the transport customer is no longer interested in distance. To the contrary, distance is a very important parameter in the cost and time functions (2) and (3). As a matter of fact, it is the main factor causing the shift of g_{ij}^{k} and h_{ij}^{k} between relations. This suggests a two-stage-least-squares solution.

1) In the first stage we regress transport costs upon distance, which is an exogeneous factor in the time- and cost equations;
2) In the second stage we apply the gravity model, the original transport costs and times having been replaced by their regression estimate upon distance.

When we use real transport costs and times in a gravity model and when our object is estimation of the demand function, it is safe to use the two-stage-procedure, outlined above. However, we have done some experiments with Belgian transport data and estimated some gravity models, first with the original cost data and second with their first stage estimate. We have to admit that the difference between parameter estimates in both cases is highly insignificant.

This suggests that simultaneous equation bias in gravity models is negligible. But we must emphasize that this comfortable conclusion only holds for cross-section estimation. In a cross-section it is true, that the variance of transport costs is almost exclusively due to exogenous factors such as distance. Compared with this predominant influence of distance (giving in the first stage determination coefficients of 0.9 or more), the influence of the transport flows themselves is negligible. To the contrary, in a time-series analysis on one given relation, distance is constant and the variance of the transport cost can to a high extent be endogenously due to the transport volume. The problem of simultaneous equations can then be extremely important.

4.2. Intermodal competition

Suppose we apply a simple gravity model:

$$x_{ij} = O_i . D_j . \beta_o (c_{ij})^{\beta_1} (t_{ij})^{\beta_2} . e^{u_{ij}} \qquad (i,j=1...I)$$

where x_{ij} is air travel from i to j;
c_{ij} is the price of the economy ticket from i to j;
t_{ij} is the transit time from i to j; including check-in times and transit to and from the airport for the average passenger.

When on a given relation ij the tariff increases, transport demand on that relation decreases with elasticity β_1^{ij}. According to our theoretical derivation of gravity models, β_1 is to be regarded as the mean of the elasticities β_1^{ij} on the various relations in our cross-section. Analogously, the estimate of β_2 could be interpreted as the mean of transit-time elasticities β_2^{ij} on the relations in the cross-section.

These interpretations are completely erroneous in the present example! We know that the disturbance term u_{ij} contains all the transport costs and -times which are present in the general demand function (1) but neglected in the gravity model. Among these are cost and time on relation ij itself by competing modes (rail and car). In a cross-section these variables show a high positive correlation with the explanatory variables c_{ij} and t_{ij}: long distances are expensive, both by air and by other modes. Short distances are cheap, both by air and by other modes.

As the effect of car and rail costs upon air transport demand is positive, the disturbance term u_{ij} and the explanatory variables show a positive covariance. The parameters are biased upwards: we may find elasticities β_1=-2 and β_2=-3 while in reality they may be -4 and -8!

We can infer from this, that in a gravity-type analysis of flows by a single mode, the specification should include additional explanatory variables, relating to cost and time by the competing modes. If not, our elasticity estimate contains the effect, not only of increasing time or costs by the mode under investigation, but also of associating with this a parallel increase of time and costs by competing

modes, the degree of parallelism being the same as in the cross-section. Nobody wants to estimate such a mixed effect!

Our estimation of cost and time elasticities in air transport only escapes from intermodal competition bias, if we extend the specification as follows:

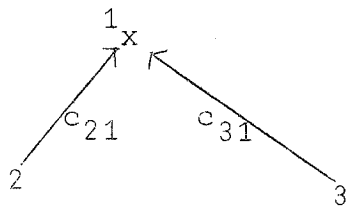
$$x_{ij} = O_i . D_j . \beta_o (c_{ij})^{\beta_1} (t_{ij})^{\beta_2} (c_{ij}^{rail})^{\beta_3} (c_{ij}^{road})^{\beta_4} (t_{ij}^{rail})^{\beta_5} (t_{ij}^{road})^{\beta_6} e^{u_{ij}}$$

where the additional variables refer to the competing modes. In this specification β_1 and β_2 precisely represent the effect of c_{ij} and t_{ij}, ceteris paribus costs and times by competing modes.

The addition of the new variables admittedly can cause multicollinearity, but multicollinearity should never be solved by leaving away important variables.

4.3. Interspatial competition

The problem of interspatial competition can be conveniently illustrated by the figure below:



Region 1 is supplied by regions 2 and 3. The quantity x_{21}, as stated by the general demand functions, depends not only on c_{21} but also on the other transport cost c_{31}. An increase of c_{31} improves the competitive position of region 2 and hence has a positive effect upon x_{21}.

However, when applying the simplest version of a gravity model, we omit transfer costs on other relations and specify

$$x_{21} = O_2 . D_1 \beta_o (c_{21})^{\beta_1} e^{u_{21}}$$
$$x_{31} = O_3 . D_1 \beta_o (c_{31})^{\beta_1} e^{u_{31}}$$
.
.

The disturbance in the first observation contains a cross price effect that could be specified as $\gamma(c_{31})$, while in the second observation it contains $\delta(c_{21})$, with cross price effects γ and $\delta > 0$.

The problem now is caused by negative correlation between the explanatory transport cost and the other transport costs, hidden in the disturbance. This is very clear from our 3-region example: in the first observation the explanatory transport cost (c_{21}) is low, while the other transport cost (c_{31}) is high. In the second observation the explanatory transport cost is high (c_{31}), while the other transport cost is low (c_{21}). When the unknown coefficients γ and δ show zero correlation with the explanatory variable c_{ij}, which is the most plausible assumption, this will result in negative covariance between u_{ij} and c_{ij}. The estimate of β_1 is biased downward. It contains the effect both of increasing the transport cost on a relation and of decreasing transport costs on other relations.

To avoid this interspatial competition bias, we can adapt the specification of the gravity model. For instance:

$$x_{ij} = O_i . D_j . \beta_o (c_{21})^{\beta_1} . (\overline{c_{.j}})^{\beta_2} . (\overline{c_{i.}})^{\beta_3} e^{u_{ij}} \qquad (i,j=1...I)$$

with $\overline{c_{.j}}$ a weighted average of the transport costs for all flows arriving in j, except the one leaving i;

$\overline{c_{i.}}$ a weighted average of the transport costs for all flows leaving i, except the one arriving in j.

This removes from the disturbance those cross price effects, which cause the negative bias shown in the example above.

In practical applications, the problem of interspatial competition bias seems to have as little importance as simultaneous equations. A lot of experiments show that the introduction of $\overline{c_{.j}}$ and $\overline{c_{i.}}$ as additional variables, does not significantly change the estimate of β_1. One of these experiments will be treated in §5.

The problem of interspatial competition bias might be merely academic, because in reality we do not face three-region problems. With many regions of origin and destination the mean values $\overline{c_{.j}}$ and $\overline{c_{i.}}$ are not substantially affected by the inclusion and exclusion of one transport cost.

4.4. Aggregation bias

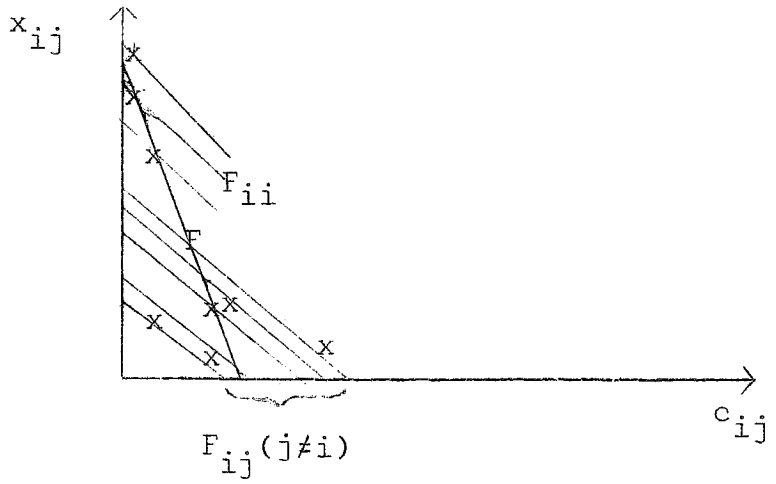
When a complete matrix of flows x_{ij} $(i,j=1...I)$ is available, nearly all analysts use the complete set of I^2 observations to estimate the parameters of the gravity model. This might give the impression that a gravity model cannot be used when some elements in the matrix are missing.

This impression , of course, is false. To the contrary, it can be desirable to eliminate some relations from the cross section even if the observations on these relations are available. When we have a specific reason to consider the demand function f_{UV} on a relation uv as entirely different from the demand functions f_{ij} on other relations, it is wise to eliminate relation uv from the regression estimation. In air transport for instance, one should run separate regressions for overseas relations and for continental relations.

In general it is also a sound practice to eliminate all the observations on the main diagonal, containing flows x_{ii} which start and end within one single region.

Even if with some abracadabra one has obtained the exact cost-, time- and flow data for these cases, the estimation of the various demand functions F_{ij} by a single F can be invalidated by the introduction

of these observations. We know that the main diagonal contains the lowest cost and time values of the whole cost and time matrices. As soon as the elasticity or the constant in demand functions F_{ii} is systematically different from the corresponding coefficients in functions F_{ij} $(j \neq i)$, we obtain an aggregation bias. This can be infered from equation (12) in §3. It is also shown by the following graph:



In this graph, we assume the demand functions on all relations to have approximately the same slope. We suppose however that for local transport (within one region) the constant of the demand function is systematically higher than the constant of interregional demand functions. This is a plausible assumption. The observations are indicated on the graph by means of x. One can notice that c_{ij} assumes the lowest values for local transport within one region (functions F_{ii}). Clearly, the slope of the function F no longer estimates the average slope of the functions F_{ij}. It contains a negative bias. The disturbance in the gravity model, containing structural differences between relations, shows negative correlation with the explanatory transport cost.

One should conclude that it is a sound rule to eliminate the main-diagonal, or in general to omit all the relation which are suspected to have structurally different demand functions in such a particular way, that the difference between coefficients is correlated with the explanatory variables.

4.5. The selection of potentials

We have been discussing four points relating to the assumption of zero-covariance between the explanatory variables and the disturbance term. We will now turn to quite another problem: the selection of suitable potentials.

In many applications of gravity models, for instance of type (4),

$$x_{ij} = O_i . D_j \beta_o . (c_{ij})^{\beta_1} . e^{u_{ij}} \qquad (i,j=1...I)$$

the potentials O_i and D_j are endogenous variables, such as total exports and imports of the regions, regional production and income or population and employment. The problem with such potentials is the fact that they are not independent from the explanatory variable c_{ij}. This has the awkward consequence that β_1 only gives the price elasticity of x_{ij}, ceteris paribus the potentials, while in reality the transport cost c_{ij} will not only have this partial effect but will also influence x_{ij} indirectly by affecting the potentials. Our estimate of β_1 neglects generated traffic (°). It neglects that influence which costs have upon flows through the change of activity (O_i and D_j) in the regions.

When we want to estimate the complete price effect of c_{ij} upon x_{ij} and not only the partial effect, potentials remaining constant, we must introduce exogenous potentials. However, exogenous variables giving the importance of a region as a transport origin or destination, are not so easy to find. Geographical surface of the region is an example, but clearly such a potential is only a vague approximation of economic importance. Potentials which merely express geographical surface will typically produce high error terms, while the role of potentials, as we have shown, is exactly the reduction of this disturbance variance.

(°) As pointed out by I.S. JONES, op.cit.

There exists a very elegant solution, which conbines the advantages both of exogenous potentials and of minimizing residual variance. It is quite incomprehensible, that this solution has been overlooked by the authors who have been applying gravity models up to now.

Instead of measuring the potentials by means of variables like regional production, surface, population or employment, we can estimate the potentials like any other coefficient in the gravity model: we only have to treat them as coefficients associated with dummy variables. For instance, we can specify:

$$\ln x_{ij} = \sum_{m=1}^{I} \rho_m . r_m + \sum_{m=1}^{I} \lambda_m . d_m + \beta_1 \ln c_{ij} + u_{ij} \qquad (,j=1...I)$$

where ρ_m and λ_m (m=1...I) are potentials, to be estimated from the cross section sample

where the summy variable r_m =1 if i=m
=o if i≠m
d_m =1 if j=m
=o if j≠m

The origin potential of a region i, ρ_i, is treated as a coefficient associated with a dummy variable that equals one for all the flows leaving region i and that equals zero for all the other flows. Analogously, the dummy associated with the destination potential λ_j is one for all the flows arriving in j and zero for the other flows.

Inevitably, as the potentials are measured by the method of least squares, this specification will contain less residual variance than specifications with potentials that are numerically specified, prior to the regression. In the case of Belgian commodity transport, this reduction of residual variance (when ρ_m and λ_m (m=1...I) are compared with regional production or total regional exports and imports) is highly significant.

It is also obvious that the coefficients ρ_m and λ_m (m=1...I) do represent exogenous potentials: they are associated with dummies which by definition do not depend on the transport costs.

§5. AN APPLICATION

For the month of october 1968, the Belgian Economic Planning Office (°) (B.E.P.) has gathered the statistics on commodity flows between 10 major regions in Belgium. We have applied several variants of the gravity model on these data and finally chosen the following specification (°°).

$$x_{ij} = \rho_i . \lambda_j . \beta_o (c_{ij})^{\beta_1} . e^{\beta_2 \delta_{ij}} . e^{u_{ij}} \qquad (i,j=1...10)$$

where x_{ij} : commodity flow from i to j (all modes);

ρ_i, λ_j : (i,j=1...10): potentials of the ten regions (to be estimated);

c_{ij} : transport cost from i to j, computed as 110 + 1,87 times the distance in kilometers between the centres of regions i and j. (The transport tariffs of the different modes were weighted with the national market share of these modes. This weighted average was regressed upon distance. Time costs were neglected as being unimportant for commodity transport demand in the Belgian distance range);

δ_{ij} : Kronecker's delta (δ_{ij}=1 for i=j and o for i≠j);

u_{ij} : disturbance term;

e : base of natural logarithms;

β_o : constant (to be estimated);

β_1 : elasticity (to be estimated);

β_2 : correction for main diagonal (to be estimated).

(°) Bureau voor Economische Programmatie, Plan 1971-1975, Produktiesectoren, Afdeling 1 : Vervoer, Brussels, 1971.

(°°) G. BLAUWENS, op.cit., pp.100 et seq.

An arbitrary fraction of the potentials ρ_i and λ_j (i,j=1...10) can be absorbed by the constant β_o. Hence we have set ρ_1 and λ_1 equal to one (omitted them from the specification), which leaves us with 21 parameters to be estimated (β_o,β_1,β_2 and 18 potentials). The main results can be summarized as follows

$\hat{\beta}_1$ = -2,4534 with standard error 0,2455
$\hat{\beta}_2$ = 1,3408 with standard error 0,2928
R^2 = 0,8939
F(20,79) = 33,29

We pretend that $\hat{\beta}_1$=-2,4534 can be considered as an estimate of the price elasticity of the demand for commodity transport: if on a relation with average elasticity transport prices increase with one percent, prices on other relations remaining constant, the demand for transport on this relation decreases with approximately 2,5 percent.

Our specification and estimation avoids some of the traps mentioned in §4:

4.1. simultaneous equations (transport costs are previously regressed upon distance);
4.2. intermodal competition (both the dependant variable and the explanatory transport cost refer to all modes);
4.4. aggregation bias (the aggregation bias, introduced by the main diagonal, is eliminated by the provision of a correction coefficient β_2);
4.5. selection of potentials (these are estimated from the sample).

There could be the problem of interspatial competition, treated in 4.3. However, adapting the specification like explained in 4.3 has a highly insignificant effect upon the elasticity estimate (°).

(°) $\hat{\beta}_1$ becoming -2,4599 with standard error 0,3398.

The effect of the additional variables furthermore is estimated with t-values inferior to one. We conclude that in our case inter-spatial competition bias is negligible.

Of course our specification presents some drawbacks. The construction of the cost index c_{ij} (the different modes being weighted rather arbitrarily) is imperfect, but this is quite a general problem occuring in the majority of econometric applications. It is not a specific problem of gravity models.

More important is the complete neglect of time costs. We hope that these are negligible in the case of commodity transport, but the inclusion of transport time as an additional explanatory variable is certainly a step for further research.

Concluding the exposition, it should be reminded that $\hat{\beta}_1 = -2,4534$ has been interpreted as the leasticity of transport demand on a single relation, when costs on this relation change and costs on other relations remain constant. It is highly erroneous to conclude that a general increase of transport costs in the whole country would have the same effect on the national transport demand. Such aggregate elasticities are quite different from the elasticity on a single relation. They cannot be estimated by means of a gravity model.