Here we consider the analysis of spatial interaction data—numbers of people, goods or services ‘flowing’ between ‘origins’ and ‘destinations’. In general, our objective is to model the pattern of flows in terms of the geographical accessibility of ‘destinations’ relative to ‘origins’, and in terms of the ‘demand’ at ‘origins’ and the ‘attractiveness’ of ‘destinations’. Such analyses are relevant in studies of health services, shopping behaviour, migration, transport planning, and in many similar areas. When flows are restricted to travel along a network, forms of mathematical optimisation and network analysis common in Operations Research (OR) also have application here, and we discuss these briefly. We also touch on location problems, since there is a reciprocal relationship between the optimal location of new public and private facilities and the understanding of observed flows to existing facilities.

9.1 Introduction

In this final chapter we turn attention to the analysis of data which, rather than being located at points or in areas, are related instead to pairs of points, or pairs of areas. Such data are usually referred to as spatial interaction data and consist of flows of some description from a set of origins to a set of destinations.

A typical example arises in a health context, where the origins may be areas of residence, which give rise to particular numbers of patients requiring treatment; these patients are then referred (or ‘flow’) to a set of destinations such as hospitals or clinics. Another example comes from retailing, where we again have a set of origin zones, generating demand for consumable goods; these demands are satisfied by retail centres and shoppers travel from origins to retail destinations. In other cases our interest might be in freight flows, population migration, or journeys-to-work. In all these cases something
(people or goods) moves from one set of places to another. But we might also consider examples where less tangible things (for example 'information' or ideas) 'flow' across space.

The primary objective in the analysis of interaction data is to understand or model the pattern of flows. If we can model such data we might then be in a position to use such a model for planning purposes. For example, if we can adequately capture the way in which patients 'interact with' hospitals we can perhaps experiment with modifications to the health care delivery system; what is likely to happen to the pattern of flow if we close a particular hospital or if we add a new service to a hospital? If we can achieve a good description of the current configuration of flows from consumer origins to retail destinations then we can speculate, with some confidence perhaps, on what might happen if a new superstore were to be opened in one of the zones, or if a new residential estate (generating further retail demand) were to be built. Both examples emphasise that our interest in spatial interaction models is motivated by the need both to understand and explain flows, and also to aid planning. As our examples make clear, the modelling and planning can take place both in a private and a public sector context.

We need to emphasise that it is not the primary concern here to understand flows at the individual level. Rather, we want to model aggregate interaction patterns. We acknowledge that there will be a myriad of reasons why individuals might choose to shop in one set of locations rather than another; similarly, there will be reasons (and health service constraints) why particular individuals visit, or are referred to, particular health centres. Our interest here lies in examining the spatial behaviour of collections of individuals, rather than individuals themselves.

How, then, do we explain such aggregate spatial interaction? Typically, we can recognise three important sets of factors:

- The spatial separation of origins from destinations
- Various characteristics which determine the volume of flow from each origin
- Various characteristics of the destinations which relate to their 'attractiveness'.

Let us briefly consider each of these sets of factors. The first relate to the way in which spatial separation or distance constrains, or impedes, movement. In general we will refer to this as the 'distance' between an origin $i$ and a destination $j$ and denote it as $d_{ij}$. At relatively large scales of geographical inquiry (such as the regional scale) this might simply be the straight-line (Euclidean) distance separating an origin from a destination. In other cases (perhaps within an urban area) it might be the travel time, cost of travel, perceived journey time, or any other sensible measure of the separation of origins and destinations. It is sometimes the case that interaction modelling is carried out in situations where origins and destinations are nodes on a transport network and flows are restricted to the arcs of that network. A sensible measure of distance from an origin node to a destination node would
then be the shortest distance between them, whether measured as arc length or travel time. In order to compute such shortest paths we may need to draw upon some of the forms of network analysis commonly used in Operations Research (OR). We shall have something more to say about these methods in the final section of this chapter. Overall, what we require is an appropriate measure of ‘distance’ which we can then use to study the extent to which spatial separation impedes movement. We shall usually expect to observe an inverse relationship between interaction and distance, known as a ‘distance decay’ or ‘distance deterrence’ relationship.

The second set of factors concerns the ability of the origin zones to generate or produce a flow. This might be measured as a single variable. For example, in a retailing context we might have data on the known, or estimated, expenditure of residents on retail goods, while in a health context we might use some measure of population size, suitably modified to allow for the fact that (depending on the particular health care application) some types of people will demand the service more than others. The important point to note is that before we begin to model the flows we may need to model ‘trip generation’. We shall expect to observe a positive relationship between the volume of flow generated from an origin and some suitable measure of the size of the origin.

A similar point may be made of the third component in the spatial interaction problem, the representation of the ‘attractiveness’ of destinations. We could measure this by variables such as the volume of floor space in the retailing context, or number of hospital beds in the health care context. Again, we shall expect to observe a positive relationship between volume of flow to the destination and the size of that destination. But there may be other factors related to the attractiveness of destinations. For example, in modelling the migration of people among Canadian provinces, or French Départements, we could measure the ‘attractiveness’ of destinations in terms of factors such as job opportunities (or unemployment rates), crime rates, climatic characteristics, and so on. Precisely what factors should be incorporated into such a model again depends on the problem; clearly, the unemployment rate is unlikely to be a determinant of retirement migration, where climate and the social landscape (involving such things as spatially varying crime rates) are likely to assume more significance.

We may not always have a matrix of flows between all possible origins and all possible destinations. In some cases we shall be interested in modelling flows from a subset of origins to a subset of destinations. For example, we might want to model the flows to one particular hospital (as in one of our case studies) or the trips made by holidaymakers to one or two particular resorts. Another example would be where we wanted to investigate the catchment area of a college or university; only one destination is the focus of interest, together with the propensity of different areas to generate students attending that institution. Both of these are examples of ‘destination specific’ problems, requiring destination specific models. Corresponding ‘origin specific’ models are available for related problems, where there may only be a subset of origins that are of interest. We return later to such models.
The basic structure of the spatial interaction problem, then, is to express the volume of flow in terms of origin, destination, and spatial separation factors. Mathematically, the situation we are considering in this chapter is one of a series of observations $y_{ij}$ ($i = 1, \ldots, m; j = 1, \ldots, n$), on random variables $Y_{ij}$, each of which corresponds to a movement of people (or cars, goods, telephone calls, and so on) between spatial locations $i$ and $j$, where these locations consist of $m$ origins and $n$ destinations. Using the notation of our previous chapters, such origins and destinations may be point locations $s_i$ or $s_j$, or alternatively they might be zones or areas $A_i$ or $A_j$. Of course, they may also be a mixture of these; for example, if we are interested in the use of recreation facilities such as swimming pools, the origins are likely to be residential areas, while the destinations will be the pools, which are fixed point locations.

In general, we will be interested in models of the form:

$$Y_{ij} = \mu_{ij} + \epsilon_{ij}$$

where, as usual, $E(Y_{ij}) = \mu_{ij}$ and $\epsilon_{ij}$ is an error about this mean. We will wish to develop suitable models for $\mu_{ij}$ involving parameters which reflect characteristics of the origins $i$ (relating to the propensity to flow from each origin); characteristics of the destinations $j$ (relating to their attractiveness); and the deterrent effect of the ‘distance’ $d_{ij}$ between origin $i$ and destination $j$.

The structure of the remainder of this chapter follows the usual lines established in earlier parts of the book. We begin by discussing a range of case studies or examples of data that lend themselves to the kinds of methods we will describe. Next, we consider the visualisation and exploration of such spatial interaction data. We then discuss more formal statistical models. Finally, we consider further and related methods which are of use in the study of spatial interaction.

### 9.2 Case studies

Throughout this chapter, we shall use a selection of data sets involving spatial interaction to illustrate various forms of analyses. We have provided copies of these data sets on disk. They include:

- Business trips made by air within Sweden
- Migration data for Dutch provinces
- Patients treated from different districts at a large London hospital
- The relative ‘attractiveness’ of different shopping centres as branch sites for a large UK financial institution.

We begin by describing these data sets in more detail; as usual, references to the sources of the data are given at the end of this chapter.

The first case study concerns journeys by air made by Swedish business travellers in 1966. Data are available on flows among 19 Swedish regions, in
each of which an airport is located (Figure 9.1). A full interaction matrix is available, showing the volume of traffic from one region to another. These data represent one way of beginning to understand the pattern of functional linkages that bind different parts of the country together. We do not have details of what such business trips entailed; whether, for example, they represent trips between members of the same, or different, organisations. But, clearly, no modern organisation can function without a flow of information, internally or externally. Whether or not the flow of information is constrained by geographical distance is, of course, a moot point. With the rapid growth in new forms of telecommunication, and the 'convergence' of places in travel time (due to technological improvements in travel) we can expect to see geographical distance shrink in importance as a factor explaining business contacts. However, such contacts in Sweden thirty years ago would indeed have been affected by distance, as well as the spatial arrangement of manufacturing and tertiary activity in the country (origin and destination factors). We should like to be able to understand, and model, the configuration of flows among these Swedish airports. In this particular data set there is a possibility that we might see the opposite relationship between flow and

![Swedish airports in 1966](image)

**Fig. 9.1** Swedish airports in 1966
distance than that we would normally expect in studies of spatial interaction, since the data relate to air travel only and longer distances may be associated with larger volumes of business air travel. Alternative forms of travel may be used for shorter distances.

The next set of data we include relate to migration among the 11 provinces of the Netherlands in 1974 (Figure 9.2). The data set also includes information on the population size of these provinces in 1974. Each item in the data set comprises the volume of flow from a particular province to the set of all provinces. This includes internal migration (within the province itself). The data thus comprise a full spatial interaction matrix.

In seeking to understand population change we obviously need information on birth and death rates, and on migration. In many developed countries it is migration that assumes particular significance when trying to understand how population change varies spatially. Since 1945 the influence of natural increase (births minus deaths) on population change has declined throughout the Netherlands. As a result, the relative contribution of internal net migration has grown. In general, what has happened in the Netherlands has been a movement of population from the highly urbanised west of the country to other areas; this ‘counter-urbanisation’ is typical of many western European countries.

Fig. 9.2 Dutch provinces
Economic factors have declined in importance, while social factors (for example, the provision of good housing, and the quality of the environment) have assumed more significance. Thus we see some important flows from provinces such as Zuid-Holland (in which the major port complex of Rotterdam is located) to Noord-Brabant further south, along with movement west from Utrecht to Gelderland. Of course, these provinces are quite large, and as so often in spatial analysis we need to be aware of scale effects; our data allow us to say nothing about quite local, short-distance moves within provinces.

These data comprise total flows and are not disaggregated by age. In other contexts we might want to focus on particular classes of mover. For example, one important category of migrants comprises retired people. Distance may well be one significant factor, but factors relating to the social and climate characteristics of the possible destinations will assume major significance. As we have said already, while we are interested here in aggregate flows there will be a host of individual factors, such as proximity to (or maybe distance from!) younger relatives that are part explanations of the individual’s reasons for a move. We simply want to make the point that the modelling of migration is a substantial topic in its own right and that there are interesting sub-groups of movers to be studied, within the general population.

Our third set of data relates to patients treated at a large London hospital. For each of a set of 14 specialities we have information on the flows of patients from 30 Health Authorities to the hospital, along with a crude measure of total population in each Authority. The data do not comprise a complete spatial interaction matrix of course; rather, they relate to a problem involving a single destination. Our interest centres on examining the way the effect of distance varies among the set of specialities. Does the drawing power, or catchment, of the hospital vary with type of speciality? Are patients with certain types of medical problem being referred to the hospital from much further away than for other medical specialities? Of course, we cannot hope to shed too much light on this problem, mainly because we lack information on other destinations. Under the current National Health Service in Britain, most hospitals (known as ‘providers’) now buy health care from Health Authorities and General Practitioners (the ‘purchasers’). In other words, hospitals must compete within an internal NHS market to provide high quality care at minimum cost. Because of this competition we cannot hope to successfully explain movement to the hospital under consideration without knowing something about competing destinations. We shall return later to the question of how such competition might be incorporated into the modelling of spatial interaction.

The application of spatial interaction analysis to health care is an important one, since it raises issues concerned with the provision of an equitable health care delivery system. What services should be provided, and where? Some medical specialities are clearly very expensive. For example, it is unrealistic to expect every district hospital to offer treatment for cancer; this is more a regional than a district level responsibility. But where should such a service be
located? Questions of equity then arise, since some patients will have to travel substantial distances, incurring costs of overcoming distance that are not only financial but also psychological. Such issues are of major concern in the developing world, where resources are limited and the need for a system of health care delivery which is both efficient and equitable is of paramount importance. We shall return later to such locational issues.

Our final case study involves data that result from the kind of models described in this chapter, rather than the interaction matrix of flows from which such models may be developed. It concerns the relative 'attractiveness' of different major shopping centres across the country as branch sites for a large UK financial institution. There are 298 shopping centres in our example, in each of which the organisation currently has a branch. There may be more than one branch in a particular centre. A competitiveness measure is included for the shopping centres which is the percentage of the number of branches of the company, in that centre, to the total number of financial organisations in that centre offering related services. The shopping centres are not the entire set of 600 such centres within which the organisation is represented, but simply the 298 most 'attractive' as determined by a measure specific to the organisation's business and developed from the kind of spatial interaction models we will be describing later in this chapter. We include the value of this 'attractiveness' measure for each of the centres. In addition we include a generally available geodemographic score for each of the centres that represents the potential drawing power of the shopping centre for general retail business rather than banking in particular.

As background we also include the volume of accounts held with the institution in question and that estimated to be held with all comparable institutions (including the institution in question) in postcode areas covering Britain. There are 118 such areas; the original study was carried out at the next lower hierarchical level of postal subdivision, and involved some 8500 zones. The figures we include were aggregated from that level.

We shall give further details later of how these data were derived and used. Suffice it to say here that the organisation was interested in 'rationalising' their branch distribution and using such results to decide to develop a new site or to close a rather unattractive one. The example is of interest in a number of general respects. First, it illustrates the fact that spatial interaction models arise in the commercial world. They are not simply 'toys' that geographers or applied statisticians like to play with. They are actually rather useful in a whole variety of retailing contexts. It is thus no surprise to find all major retailers, both in Britain and abroad, with specialist departments that undertake research into the flows of customers to retail centres, mainly with a view to predicting the likely effects of opening or closing branches. Regardless of whether these retailers are the major supermarkets, the leading motor manufacturers, or large financial organisations such as the one we are considering, they all make use of the sorts of models we shall review later.

A second reason why this particular example is of interest is that it illustrates that in some applications many of the flows we observe in practice will be zero!
Traditionally, analysis of spatial interaction data was based on rather small numbers of origins and destinations. In the larger study from which the results we include were derived, ‘flows’ of accounts were modelled between 8500 areas of residence and over 600 branches, throughout Britain. We would obviously expect vast numbers of the interactions between origins and destinations to be zero, simply because we would not expect current accounts to ‘flow’ over distances of, say, more than about 50 kilometres. Account holders in Aberdeen, in north-east Scotland, are unlikely to use branches in London! Such flows we might deem as ‘infeasible’ and the fact that they are observed not to occur does not contribute any useful information to our understanding or modelling of those that do. On the other hand, there will be some observed zero flows that are of interest to the organisation; for instance, where there is business arising in one area that does not seem to be finding its way to reasonably close centres. These ‘feasible’ zero flows do contribute to the understanding of the pattern of existing flows; an observation of zero is still an observation. They need to be identified and included in the data upon which models are based.

These are but a few case studies. They are sufficient, however, to allow us to reiterate that spatial interaction problems arise in a variety of contexts, both social and economic, and within both the commercial and public facility arenas.

9.3 Visualising and exploring spatial interaction data

Given an \((m \times n)\) matrix of observed flows, \(y_{ij}\), how might we represent this graphically? Clearly, we run the risk of producing rather cluttered maps if we endeavour to display all \(m \times n\) flows! Where interest centres on origin specific problems (or, conversely, destination specific situations) it is quite straightforward to map the flows using line symbols, where the width of the line is proportional to the volume of flow, using a sensible unit width. This is similar in spirit to the mapping of point or area data using proportional symbols that we encountered in earlier chapters. Arrowheads may be added to show the direction of flow. The American cartographer Waldo Tobler has created computer software for displaying such data (Figure 9.3). Note that the example shows net flows (here relating to migration among the coterminous states in the US) rather than the entire interaction matrix. Display is clearly more problematic when dealing with an entire interaction matrix, though Tobler’s software also includes ways of doing this.

Before we examine more formal models for spatial interaction we should also look briefly at some exploratory techniques for representing such data. Some of these involve drawing simple scatter diagrams, others involve some exploratory statistical techniques linked to visual representations.

One simple way of exploring the data is to plot the volume of interaction from (or to) a particular place against the size of the origin (or destination),
and against the chosen measure of spatial separation. As we have already seen, we can expect to observe a positive relationship between the level of interaction and ‘size’ of origin or destination. Similarly, we can anticipate a negative relationship between interaction and distance. A simple approach is to plot the proportions of flow, $y_{ij}/\sum_j y_{ij}$, out of a particular origin $i$ to different destinations $j$, against $d_{ij}$. We could do the same with the proportions, $y_{ij}/\sum_i y_{ij}$, into a destination $j$ from different origins $i$. Alternatively, we can correct for size effects by dividing $y_{ij}$ by independent measures of origin or destination ‘size’. Unless we do this, plots of interaction against distance, without simultaneously accounting for ‘size’ effects, are unlikely to reveal the true relationship because of the effect on interaction of large centres some distance from the origin or destination.

For example, with migration in the Dutch provinces, an obvious candidate for such a ‘size’ effect is the population of a destination. So we might try plotting $y_{ij}/P_j$ against $d_{ij}$ for selected origins $i$, where $P_j$ is the population of destination $j$. If we do this for the number of migrants leaving Drenthe, in the north-east of Holland, for other destinations (one of which is Drenthe itself) we see that there is a strong distance decay effect as shown in Figure 9.4a. The relationship is far from linear. Taking logarithms of both variables generates a plot that is nearly linear, shown in Figure 9.4b, suggesting a power relationship with distance. An alternative would be to take the logarithm not of distance, but only of the other variable, exploring for an exponential relationship. Note that here we have taken the intra-zone distance (Drenthe to Drenthe) to be half the distance to the nearest neighbouring zone, to avoid taking the logarithm of zero. From this graph we can pick out zones whose interaction appears to deviate from the general trend. In this case, Groningen, to the north of Drenthe, seems to be rather more attractive as a destination than its distance might predict. This may genuinely be the case; on the other hand, bearing in mind that the zones are rather large, and our distance measure is crude, it may
simply be an artifact of the scale and nature of our zoning system. This seems a particularly appropriate place at which to point out that results from spatial analysis in an interaction context are every bit as affected by the nature of the zoning system as they are when analysing the spatial arrangement of values in the areas themselves.

What other procedures are available to us to explore the structure of interaction matrices? The methods we describe in the remainder of this section assume that the set of origins is the same as the set of destinations; thus we are dealing with a square interaction matrix. This is not always the case, as seen in our earlier case study involving 'flows' of bank accounts. Note that when it is the case, the interaction matrix is not necessarily symmetric. Indeed it is exactly the question of whether $y_{ij}$ is the same as $y_{ji}$ that some of these methods are designed to highlight and explore.

Some exploratory tools seek to uncover evidence of hierarchical structure in interaction data. For example, if we are dealing with a flow matrix in which the origins and destinations are urban areas, clearly some of these will be 'located' at higher positions in the urban hierarchy than others. Some centres will therefore come to 'dominate' others; where interaction is strongly constrained by geographical distance this dominance may take the form of a single centre acting as a 'magnet' for the surrounding area. More generally, however, a centre will dominate others, not in a simple geographical way but in terms of functional linkages. We can use the flow data in an interaction matrix to reveal the patterns of dominance. A simple way to do this graphically is to represent each place as a point on a topological map and to link $i$ to $j$ with a directed arc if the largest flow from $i$ is to $j$ and if $j$ is larger than $i$ (where size can be measured by the column totals in the interaction matrix). This procedure can be made more sophisticated, but it yields some evidence of hierarchical structure in the form of a directed graph.
Waldo Tobler, mentioned earlier in connection with computer mapping of flow data, has also devised some highly imaginative ways of exploring interaction data. From the asymmetric matrix of interactions we may compute:

\[ c_{ij} = \frac{(y_{ij} - y_{ji})}{(y_{ij} + y_{ji})} \]

where \( c_{ij} \) may be interpreted as a ‘current’ that ‘aids’ flow from \( i \) to \( j \) if \( y_{ij} > y_{ji} \). This current may be displayed as a vector drawn from \( i \) to \( j \) of length \( c_{ij}/2 \) (note that \( c_{ij} = -c_{ji} \)). Repeating this for all flows to and from \( i \) yields a cluster of vectors at \( i \), the resultant total vector (summing both directions and magnitude) being interpreted as a ‘wind’ at \( i \). Repeating this for all locations gives a vector field which Tobler refers to as ‘winds of influence’; the vector field may be interpolated to a regular grid if there are sufficient origins/destinations. Tobler develops the meteorological analogy further by deriving, from the vector map of winds, a ‘pressure field’ that is conceived of as giving rise to the winds. Interestingly, Tobler has applied his method to the Swedish data on air travel, deriving a map that shows a net flow towards Kiruna in the north of the country, and Göteborg and Malmö to the south (Figure 9.5).

An interaction matrix may be thought of as a kind of ‘similarity’ matrix, so we can also consider using some of the multivariate methods discussed in Chapter 6 designed to explore structure in similarity or dissimilarity matrices (one is just minus the other). The technique of multidimensional scaling described there is sometimes appropriate to analysing interaction patterns. Recall that this technique is concerned with deriving a configuration of a set of locations in a space of minimum dimensionality, in such a way that the Euclidean distance between them in this space approximates as closely as possible a given dissimilarity matrix. In this case we could base our dissimilarity matrix on minus the interaction matrix of flows. As a result, we might seek a representation of the areas such that zones between which a lot of interaction occurs are located close together in ‘interaction space’, while zones that interact little are distant. This idea has sometimes proved useful in the exploration of interaction data.

### 9.4 Modelling spatial interaction data

We move on now to consider more formal models for spatial interaction data. As mentioned in the introduction we are interested in modelling observed flows according to a statistical spatial interaction model of the general form:

\[ Y_{ij} = \mu_{ij} + \epsilon_{ij} \]
Fig. 9.5 ‘Winds of influence’ in Swedish air travel

where $E(Y_{ij}) = \mu_{ij}$ and $\epsilon_{ij}$ is a residual error term. Note that this residual term relates to a pair of locations; we model the individual flows between $i$ and $j$, not simply the total flow from $i$ or to $j$.

What specific mathematical form for $\mu_{ij}$ can be adopted that will reflect, in a suitable parametric way, the various origin, destination and distance deterrence effects that we referred to earlier?
9.4.1 Basic spatial interaction or gravity models

One of the most common classes of models for $\mu_{ij}$, which simultaneously incorporate the effect of origin and destination characteristics as well as distance, are known as gravity models. The term arises because of their analogy with Newton’s law of gravity, where force of attraction is proportional to the product of the masses of the two bodies involved and inversely proportional to the square of the distance between them. In our case ‘flow’ is analogous to force and the origin and destination effects would be analogous to the masses of the two bodies. Such models were originally used simply because they seemed a sensible and convenient mathematical way to represent spatial interaction. Unfortunately, a simple model that predicts interaction as proportional to the product of the origin and destination ‘size’ and inversely proportional to distance does not always turn out to be adequate. This is because it may generate estimates of interaction that, when summed over rows or columns, give results that are inconsistent with the known number of flows leaving an origin or arriving at a destination and we may wish our model to ‘reproduce’ these totals exactly. Nevertheless, it can be demonstrated that the simple Newtonian model is one of a wider class of models with a mathematical form that can be theoretically justified on the basis of entropy maximisation. This provides a more solid foundation for the widespread use of such models.

The argument behind the entropy maximisation approach is as follows. Since it is the mean or expected flow $\mu_{ij}$ that we are attempting to model, we consider how we might theoretically expect a set of such mean flows to arise. We start with a situation involving a total number, $N = \sum \sum \mu_{ij}$, of such flows and think of this as a system of $N$ ‘average mobile individuals’ who are free to arrange themselves as they wish in order to constitute any particular set of average flows, $\mu_{ij}$, between origins $i = 1, \ldots, m$, and destinations $j = 1, \ldots, n$, but we insist that they do so subject to three overall constraints that we consider are fixed in our system:

1. A certain total number, $a_i$, have to flow from each origin i.e.:

$$\sum_j \mu_{ij} = a_i$$

2. A certain total number, $b_j$, have to flow to each destination i.e.:

$$\sum_i \mu_{ij} = b_j$$

3. The total ‘cost’ of travel in the whole system, $c$, is fixed i.e:

$$\sum_i \sum_j d_{ij} \mu_{ij} = c$$

Recall that we said earlier that we were not interested in the individual mover. We would say that we are ignorant of the details of such a move. Here, however, we are saying that we will have certain information available to us
about aggregate flows on a system-wide basis; we know how many trip-makers leave each origin, how many arrive at destinations, and we have added a further piece of information, about the total system-wide ‘cost’ of travel. We seek a solution as to how we expect mean flows to arrange themselves within the constraints imposed by this global information. We do this by considering the question:

‘out of all possible arrangements of the N individuals into sets of flows $\mu_{ij}$ that would satisfy these constraints, which is the most likely, or the most probable?’

Using basic ideas of permutations and combinations we can write down the number of different ways that the total number of ‘average individuals’, $N = \sum_i \sum_j \mu_{ij}$, can be assigned to particular flows in such a way that we obtain $\mu_{ij}$ of them in the flow from $i$ to $j$. We get the expression:

$$\frac{\left( \sum_i \sum_j \mu_{ij} \right)!}{\prod_{i,j} \mu_{ij}!}$$

This quantity is often referred to as the entropy function. Finding the most likely $\mu_{ij}$ out of all possible sets of flows that would satisfy the three constraints outlined earlier, is essentially equivalent to choosing values for the $\mu_{ij}$ in such a way as to maximise this entropy function subject to those constraints.

We do not give mathematical details of the maximisation here. Some of the references we give at the end of the chapter give detailed explanations of what is involved. Briefly, the logarithm of the entropy function is maximised, using Lagrange multipliers to take account of the three constraints. The result obtained is that any choice of $\mu_{ij}$ that will maximise the entropy function must satisfy the general equation:

$$- \log (\mu_{ij}) - \lambda_i^{(o)} - \lambda_j^{(d)} - \lambda d_{ij} = 0$$

where $\lambda_i^{(o)}$ are the Lagrange multipliers used to take account of the origin constraints; $\lambda_j^{(d)}$ those to take account of the destination constraints; and $\lambda$ that to take account of the total cost constraint.

Entropy maximisation therefore leads to a model for $\mu_{ij}$ of the general form:

$$\mu_{ij} = \alpha_i \beta_j e^{\gamma d_{ij}}$$

where $\alpha_i = e^{-\lambda_i^{(o)}}$, $\beta_j = e^{-\lambda_j^{(d)}}$ and $\gamma = -\lambda$ simply re-express the Lagrange multipliers in a more convenient form.

This is referred to as the general gravity or spatial interaction model. In this model, $\alpha_i$ are interpreted as a set of parameters which characterise the propensity of each origin to generate flows, $\beta_j$ a set of parameters which characterise the attractiveness of each destination and $\gamma$ a distance deterrence effect.

351
As stated previously the general form of the gravity model was used extensively to model observed flows purely on intuitive grounds before it was theoretically justified as the general solution to an entropy maximisation problem as described above. Note that if $d_{ij}$ is taken as the logarithm of distance then the model is of the form:

$$\mu_{ij} = \alpha_i \beta_j d_{ij}^\gamma$$

and the model in this form resembles the Newtonian law of gravitation in the special case $\gamma = -2$.

### 9.4.2 Estimating the parameters of gravity models

Having derived a particular functional form for $\mu_{ij}$ which has some theoretical justification, we can now return to the problem of modelling a set of observed flows $y_{ij}$. Recall that we think of these as observations on random variables $Y_{ij}$ with mean value $\mu_{ij}$ and the general model proposed earlier was:

$$Y_{ij} = \mu_{ij} + \epsilon_{ij}$$

Incorporating our general gravity model for $\mu_{ij}$, this now becomes:

$$Y_{ij} = \alpha_i \beta_j e^{\gamma d_{ij}} + \epsilon_{ij}$$

From a statistical point of view, fitting this kind of gravity model to observed data is a question of estimating the unknown parameters $\alpha_i$, $\beta_j$ and $\gamma$. As a first approach it is tempting to take logarithms of this model and write it in a linear form as:

$$\log Y_{ij} = \alpha'_i + \beta'_j + \gamma d_{ij} + \epsilon'_{ij}$$

and then proceed to estimate parameters by using an ordinary least squares regression of the observations $y_{ij}$ on $d_{ij}$ and a set of ‘dummy’ (0,1) variables to represent the origin parameters, $\alpha_i$, and the destination parameters, $\beta_j$. However, such an approach suffers from two drawbacks.

Firstly, there is no guarantee that flows predicted from the model will necessarily have the property that, when summed by rows or columns, they agree with either the total number observed to flow from each of the origins, or with the total number observed to flow to each of the destinations. Similarly the predicted ‘total cost of travel’ in the system will not necessarily correspond to that observed. Agreement of these totals would seem to be a desirable property of any ‘sensible’ estimates for the parameters $\alpha_i$, $\beta_j$ and $\gamma$, since they are the ‘system-wide’ fixed constraints which we used to justify the functional form for $\mu_{ij}$ used in the model. If we wished to enforce this, then the log-linear regression model given above would need to be modified to incorporate explicitly the required constraints. This would turn it into a non-linear model, and the estimation of parameters then becomes very much more difficult.
Secondly, estimating parameters by the ordinary log-linear regression model given above would only be justified statistically if we believed that flows \( Y_{ij} \) were independent and log-normally distributed about their mean value with a constant variance. Such an assumption is patently not valid since flows are discrete counts whose variance is very likely to be proportional to their mean value. Least squares assumptions ignore the true integer nature of the flows and approximate a discrete-valued process by an almost certainly misrepresentative continuous distribution. As a result, ordinary least squares regression estimates and their standard errors can be seriously distorted.

For both of the above reasons, gravity models cannot be fitted by an unconstrained log-linear regression. Maximum likelihood estimation of parameters under more realistic distributional assumptions is generally considered a more suitable approach. The most common assumption is that \( Y_{ij} \) follow independent Poisson distributions with expected values \( \mu_{ij} = \alpha_i \beta_j e^{\gamma d_{ij}} \). Such assumptions are also open to question, since flows are not strictly independent and, furthermore, a Poisson distribution may not adequately reflect the degree of variation present in many real data sets, since in many applications individuals may tend to ’flow’ in groups rather than as individuals. Nevertheless these assumptions are generally regarded as leading to ‘reasonable’ parameter estimates, at least in the first instance. The alternative would be to use some distributional assumption more able to reflect ‘over dispersion’; computational problems involved in parameter estimation can then become considerable. We give some references to work on these sorts of problem at the end of this chapter.

Under the assumption that the \( Y_{ij} \) are independent Poisson random variables with means \( \mu_{ij} = \alpha_i \beta_j e^{\gamma d_{ij}} \), the basic gravity model can be treated simply as a particular case of a generalised linear model with a logarithmic link and a Poisson error. We discussed such models in some detail in Chapter 8. Recall that maximum likelihood estimates of parameters are derived by means of iterative re-weighted least squares (IRLS), which is implemented as a standard procedure in many statistical packages such as GLIM.

But what of the requirement, mentioned earlier, that a desirable property of parameter estimates should be that predicted flows from the model, when summed over origins and destinations, should agree with observed totals leaving origins and arriving at destinations and that the total predicted ‘cost of travel’ should agree with that observed? A convenient property of the Poisson assumption is that resulting maximum likelihood parameter estimates \( \hat{\alpha}_i, \hat{\beta}_j \) and \( \hat{\gamma} \) automatically guarantee that fitted flows \( \hat{Y}_{ij} = \hat{\alpha}_i \hat{\beta}_j e^{\hat{\gamma} d_{ij}} \) satisfy relationships that are entirely consistent with the desirable origin, destination and total cost of travel constraints. This is not too difficult to see since the log likelihood for the general gravity model under the assumption that \( Y_{ij} \) are independent Poisson random variables with means \( \mu_{ij} = \alpha_i \beta_j e^{\gamma d_{ij}} \), is effectively:

\[
\sum_{i=1}^{m} \sum_{j=1}^{n} (y_{ij} (\log \alpha_i + \log \beta_j + \gamma d_{ij}) - \alpha_i \beta_j e^{\gamma d_{ij}})
\]
The analysis of spatial interaction data

To maximise this we need to differentiate with respect to each of the parameters in turn and set the resulting derivatives to zero. The maximum likelihood parameter estimates \( \hat{\alpha}_i, \hat{\beta}_j \) and \( \hat{\gamma} \) must therefore satisfy any equations that result. Setting the derivatives, with respect to each \( \alpha_i \), to zero leads to the estimates having to satisfy:

\[
\sum_{j=1}^{n} \left( y_{ij} - \hat{\alpha}_i \hat{\beta}_j e^{\hat{\gamma}d_{ij}} \right) = 0
\]

In other words, if we let \( \hat{y}_{ij} = \hat{\alpha}_i \hat{\beta}_j e^{\hat{\gamma}d_{ij}} \) be the predicted flow from \( i \) to \( j \) then:

\[
\sum_{j=1}^{n} \hat{y}_{ij} = \sum_{j=1}^{n} y_{ij}
\]

The same process using the derivatives with respect to each \( \beta_j \) gives:

\[
\sum_{i=1}^{m} \hat{y}_{ij} = \sum_{i=1}^{m} y_{ij}
\]

and the derivative with respect to \( \gamma \) gives:

\[
\sum_{i=1}^{m} \sum_{j=1}^{n} \hat{y}_{ij}d_{ij} = \sum_{i=1}^{m} \sum_{j=1}^{n} y_{ij}d_{ij}
\]

These are exactly the constraints imposed on theoretical mean flows discussed in relation to entropy maximisation earlier where:

\[
\sum_{j} y_{ij} = a_i
\]

\[
\sum_{i} y_{ij} = b_j
\]

\[
\sum_{i} \sum_{j} y_{ij}d_{ij} = c
\]

Thus, maximum likelihood parameter estimates of the general gravity model under the assumption of independent Poisson flows automatically ensure predicted flows that reproduce the observed total flow from each origin and to each destination as well as the observed ‘total cost of travel’.

This is a convenient property of the Poisson assumption, since it avoids the need to modify standard maximum likelihood parameter estimation to incorporate explicit constraints on predicted flows. It also suggests a simple iterative algorithm for fitting Poisson gravity models, which may be used as an alternative to direct maximisation of the likelihood and will arrive at the same parameter estimates. This may be particularly useful when the number of
origins or destinations is large, since in that case treating a Poisson gravity model as a particular case of a standard general linear model and using a computer package like GLIM to fit it, may present problems. Statistical packages such as GLIM may have difficulty handling very large data matrices and the general IRLS algorithm which they use to find maximum likelihood estimates is not particularly computationally efficient for this particular problem when large numbers of parameters are involved.

This alternative algorithm proceeds as shown in Table 9.1. The algorithm is repeated until convergence to give a final set of parameter estimates \( \hat{\alpha}_i \), \( \hat{\beta}_j \) and \( \hat{\gamma} \), which are entirely equivalent to those that would be obtained from general maximum likelihood parameter estimation. However, no standard errors for the estimates are naturally produced from this algorithm. They may be obtained from the usual large sample result involving the inverse of the second derivative of the likelihood function evaluated at final parameter estimates. This involves further computational steps, details of which we do not give here.

### Table 9.1  Iterative estimation of parameters for doubly constrained model

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Step 0</strong></td>
<td>Set ( s = 0 ), where ( s ) denotes 'step', and take initial estimates ( \hat{\alpha}^{(0)}_i = 1 ) (( i = 1, \ldots, m )) and ( \hat{\beta}^{(0)}_j = 1 ) (( j = 1, \ldots, n ))</td>
</tr>
<tr>
<td><strong>Step 1</strong></td>
<td>Maximise the Poisson log likelihood conditional on the current values of ( \hat{\alpha}^{(s)}_i ) and ( \hat{\beta}^{(s)}_j ) with respect to the single parameter ( \gamma ) to obtain a current estimate ( \hat{\gamma}^{(s)} ). Since there is a single parameter involved this is relatively easy and may be done routinely using a simple numerical optimisation method such as Newton–Raphson.</td>
</tr>
</tbody>
</table>
| **Step 2** | Since it is known that the final parameter estimates must produce predicted flows \( \hat{\alpha}_i \hat{\beta}_j e^{\gamma d_{ij}} \) which satisfy the origin and destination constraints, iteratively proportionately adjust the current \( \hat{\alpha}^{(s)}_i \) and \( \hat{\beta}^{(s)}_j \) so that such constraints are satisfied for the current predicted flows. That is, adjust \( \hat{\alpha}^{(s)}_i \) according to:  
\[
\hat{\alpha}^{(s)}_i = \frac{\sum_j y_{ij}}{\sum_j \hat{\beta}^{(s)}_j e^{\gamma d_{ij}}},
\]  
and then \( \hat{\beta}^{(s)}_j \) according to:  
\[
\hat{\beta}^{(s)}_j = \frac{\sum_i y_{ij}}{\sum_i \hat{\alpha}^{(s)}_i e^{\gamma d_{ij}}},
\]  
and keep repeating these calculations until the values of \( \hat{\alpha}^{(s)}_i \) and \( \hat{\beta}^{(s)}_j \) stabilise to final values which are then designated \( \hat{\alpha}^{(s+1)}_i \) and \( \hat{\beta}^{(s+1)}_j \). This is known as the balancing step. |
| **Step 3** | Set \( s = s + 1 \) and return to Step 1. |
Before proceeding further it may be useful to illustrate some of these ideas with reference to one of the case studies described earlier. The results of fitting a gravity model of the type described above to the data on migration between Dutch provinces are given in Table 9.2. Here we have used flows in thousands of individuals and a crude measure of separation based on the distance in kilometres between the ‘centres’ of provinces. For example, the predicted flow (000s) from Drenthe to Groningen, which has a separation of 42.9 km, is given by:

\[
\hat{y}_{14} = 10.45 \times 1.065 \times e^{-0.0241 \times 42.9} = 3.96
\]

The observed flow is 3.89, so the model over predicts slightly in this case. However, this model will predict total flows to Groningen from all provinces which correspond to observed total flows; similarly with total flows out of Groningen to all provinces. This will also apply for any other of the eleven provinces.

Note that values \(\hat{\alpha}_i\) or \(\hat{\beta}_j\) from such models are arbitrary in the sense that one could multiply each \(\hat{\beta}_j\) by any constant and divide each \(\hat{\alpha}_i\) by the same factor and still obtain the same predicted flows. Nevertheless, relative comparisons of \(\hat{\alpha}_i\) between origins or \(\hat{\beta}_j\) between destinations may be made. Their interpretation is, respectively, the relative ability of origins to generate flows, or destinations to attract flows, after the deterrence effect of distance has been allowed for.

The general form of the gravity model that we have discussed so far and used in the above example is often referred to as a **doubly constrained model**. We have allowed \(\alpha_i\) and \(\beta_j\) to be separate parameters for each origin and each destination and have not attempted to ‘explain’ their values, simply to estimate them in such a way that predicted flows reproduce exactly the total observed

<table>
<thead>
<tr>
<th>Province</th>
<th>(\hat{\alpha}_i)</th>
<th>(\hat{\beta}_j)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drenthe</td>
<td>10.45</td>
<td>0.658</td>
</tr>
<tr>
<td>Friesland</td>
<td>16.72</td>
<td>0.845</td>
</tr>
<tr>
<td>Gelderland</td>
<td>31.11</td>
<td>1.448</td>
</tr>
<tr>
<td>Groningen</td>
<td>20.27</td>
<td>1.065</td>
</tr>
<tr>
<td>Limburg</td>
<td>30.00</td>
<td>1.115</td>
</tr>
<tr>
<td>Noord-Brabant</td>
<td>33.46</td>
<td>1.482</td>
</tr>
<tr>
<td>Noord-Holland</td>
<td>51.54</td>
<td>1.440</td>
</tr>
<tr>
<td>Overijssel</td>
<td>20.01</td>
<td>0.944</td>
</tr>
<tr>
<td>Utrecht</td>
<td>20.70</td>
<td>0.770</td>
</tr>
<tr>
<td>Zeeland</td>
<td>13.53</td>
<td>0.558</td>
</tr>
<tr>
<td>Zuid-Holland</td>
<td>59.92</td>
<td>1.615</td>
</tr>
</tbody>
</table>

\[\hat{y} = -0.0241 \quad (0.0009)\]
flow from each origin and to each destination; hence the name ‘doubly constrained’.

In a sense such a model is purely descriptive rather than offering an explanation for the observed flows, and in the next section we will go on to discuss modifications to the model which allow us to incorporate more explanation into \( \alpha_i \) and \( \beta_j \). However, such doubly constrained models do offer one element of ‘explanation’ relating to observed flows. The distance parameter \( \gamma \) quantifies the relative ability of distance (or, more generally, cost of travel) to deter spatial interaction, and this is useful information. As mentioned above, the model also allows relative comparison of the ‘generating potential’ of origins, or the ‘pull’ of destinations, after due allowance has been made for distance effects.

We can also use this model to predict what flows will result if certain changes are made in the system. For example, if we wished to see what would happen if twice as many people migrated from Drenthe as at present, but only half as many from Groningen, then we would simply double the \( \alpha_i \) estimate corresponding to Drenthe, halve it for Groningen and then use the model to predict resulting flows. Similarly, if the distance measures used in the model relate to travel time or travel costs, then we could use the model to predict what would happen if the travel time between certain origins and destinations were to change. Note that when we use the model in these ways we are assuming that the same origins and destinations are always involved and we know in advance how many will travel in total from each origin and arrive at each destination. We have no way of introducing hypothetical origins or destinations into the system, because we would not know how to assign their \( \alpha_i \) or \( \beta_j \) values. Neither can we ask what would happen if the population grows faster in one of the Dutch provinces that in others, because no relationship between population and the numbers migrating is built into the model.

For these reasons the usefulness of doubly constrained models is somewhat limited and we shall take up alternative models in the next section. However, their predictions are useful in some areas. For example, doubly constrained models are often used in transport planning, in modelling the journey to work. Here, we might argue that we know, or can predict by other means, the numbers leaving origin zones and the numbers employed at destinations. What we seek are good predictions of flows from \( i \) to \( j \), given these numbers. This would allow us to assess whether the transport network was capable of dealing with existing flows. It might permit us to identify bottlenecks on the road network, for example. The model could also be used to simulate the possible journey-to-work flows if new roads were built, if one-way systems were devised, or if some routes were banned to motor vehicles. This is done by modifying particular elements in the distance, or cost, matrix. For example, such models fitted to our Swedish airline flow data might help to predict where there might be demand for new airline routes and what pricing structure should be employed.

We see that these models, while inevitably simplified, can begin to be used in real planning contexts, even to the extent of informing transport policy.
Further, while we do not consider such extensions, we note that the model can be \textit{disaggregated} to take into account different modes of transport. The introduction of such \textit{modal split} into such models allows us to separate out car users from cyclists, bus travellers, and so on.

\section*{9.4.3 Variations of the gravity model}

We referred earlier to the general gravity model discussed in the last section as a \textit{doubly constrained model}. To recap, we allowed it to have a separate parameter for each origin and each destination, enabling the fitted model to reproduce exactly the total observed flow from each origin and to each destination.

We repeat that such a \textit{doubly constrained model} is purely descriptive rather than offering much explanation for the observed flows. The distance parameter does quantify the relative ability of distance (or, more generally, cost of travel) to deter spatial interaction, but the origin and destination parameters offer no real explanation of what causes individuals to flow, or why one destination is apparently more attractive than another, when its relative distance from origins is allowed for. They are only of use in making relative comparisons between origins or destinations, not in explaining differences.

It is often the case that one has further attributes measured at either origins or destinations or both, and one wishes to use this additional information to model either the number of individuals that will flow from any origin, or the relative attractiveness of destinations. We saw this in relation to some of the case studies described earlier.

One way to approach this question would be as a two stage process. A doubly constrained model is fitted to the observed flows and, as a separate exercise, the origin or destination parameter estimates, \(\hat{\alpha}_i\) and \(\hat{\beta}_j\), so obtained, are in turn modelled in terms of the additional covariate information present at either origins or destinations. In preliminary analysis this may well be a useful way to establish what covariates are potentially useful and what form of relationships should be used to incorporate them into the flow model.

Indeed, an examination of how well an interaction model fits individual flows between particular geographical areas should also be emphasised. Studies of selected model residuals (differences between predicted and observed flows) can be performed to examine local effects not accounted for in the model and to investigate possible explanations for the attractiveness estimates \(\hat{\beta}_j\) and the source potential estimates \(\hat{\alpha}_i\) in more detail. It should be appreciated that the basic spatial interaction model has a very simple structure and undoubtedly masks many complexities in the observed flows. Several characteristics of localities can act as actual or perceived barriers to travel or otherwise influence flows in ways not accounted for by the basic model.

These kinds of approaches are relevant to the case study referred to earlier, concerning the 'attractiveness' of different shopping centres as sites for branches of a large British financial institution. As mentioned, our data are
only some of the results of the original modelling exercise. The original study was motivated by wishing to derive, for each major shopping centre throughout the British Isles, a measure specifically oriented to their ‘attractiveness’ for siting branches of the organisation offering certain kinds of banking services. The flow data comprised the volume of accounts held in each existing branch by people residing in different areas of the country, broken down into some 8500 postal zones. This large and very sparse interaction matrix was used to fit a doubly constrained gravity model to the observed ‘flows’ of accounts, resulting in relative ‘attractiveness’ estimates, $\hat{\beta}_j$, for the 600 shopping centres containing branches. The idea was then to model these ‘attractiveness’ estimates in terms of exogenous measures, such as size of centre, level of competition and so on. If reasonable models could be found to explain differences in ‘attractiveness’, then it would be possible to extrapolate meaningful estimates of the ‘attractiveness’, for the organisation’s existing customers, of the remaining 1600 shopping centres where the organisation was currently not represented. This would enable them to rationalise their distribution of branches, perhaps closing some down in order to open others in new centres. Further, if the exogenous variables found to explain centre attractiveness were such that they could be adjusted for competitors, it would also be possible to predict the attractiveness of each of the 2200 centres to customers of competitors. Given estimates of competitor customer demand and the location of competitor outlets, it might then be possible to choose optimal sites for expansion. We do not give any details here of the results of such an exercise.

Studies such as this are clearly aided by being able to look in detail at various aspects of the model results, in particular to relate them to local geography and infrastructure. Until recently this has been somewhat difficult in practice, particularly in modelling exercises involving large interaction matrices. However, even relatively simple mapping packages provide an excellent vehicle to explore and evaluate interaction models in more detail at a local level. More powerful GIS systems have an even greater potential in this area.

These sorts of analyses of predicted values and origin and destination parameter estimates arising from a doubly constrained model, may thus result in suggesting covariates which might be included into the flow model to explain, rather than simply describe, origin and destination effects. The general form of the doubly constrained model can be naturally adapted to incorporate such covariates. An entire ‘family’ of spatial interaction models can be conceived, each involving different ways of modelling $\alpha_i$ and $\beta_j$ and resulting in different constraints on predicted flows. Having looked in detail at the doubly constrained member of this family, let us now briefly consider others.

In the origin constrained model the destination parameters, $\beta_j$, of the doubly constrained model are replaced by some (usually simple) function of $p$ observed covariates $x_{ij}^{(d)} = (x_{ij1}^{(d)}, \ldots, x_{ijp}^{(d)})$ at each of the destinations $j$. The idea is that the ‘attractiveness’ of destinations may be characterised in terms of a few measurable attributes of those locations. It is usual to include the covariates in a log-linear way so that the model becomes:
\[ Y_{ij} = \alpha_i e^{v(x_j^{(d)}; \theta)} + \gamma_{ij} + \epsilon_{ij} \]

where \( v(x_j^{(d)}; \theta) \) is some function (usually linear) of the vector of destination characteristics \( x_j^{(d)} \), involving a vector of associated parameters \( \theta \). In such a model the total predicted flows to each destination are no longer constrained to equal total observed flows to that destination. However the observed total flow from any origin remains exactly reproduced by the model, hence the terminology origin constrained (sometimes known as production constrained).

One important area of application of such a model arises in retailing. To take a simple example, we might represent the attractiveness of a destination (which, in a modelling context, would be a zone containing retail outlets) in terms of the volume of, say, floor space. In this case the vector of destination characteristics and parameters would be replaced by only one such characteristic. In the origin constrained model, predicted flows to a particular destination are not constrained to always equal the observed total flow at that destination, so the model allows us to obtain predictions of total flows, or retail expenditure, at the shopping zones, given alternative configurations of floor space in these zones. The origin constrained model therefore allows us to predict total flow at destinations as a function of floor space, given a fixed pattern of demand at the origins. This may be particularly useful in predicting shopping centre usage, and in aiding the evaluation of existing, or new, sites. For example, in a competitive environment we may conceive of using such a model to allocate floor space at retail outlets belonging to our organisation or providing floor space at new outlets in such a way as to maximise revenue to us as opposed to our competitors. We shall return to these kinds of optimisation problems in the final section of this chapter.

Under the Poisson assumption for \( Y_{ij} \) the origin constrained model can again be fitted by maximum likelihood to obtain estimates \( \hat{\alpha}_i, \hat{\theta} \) and \( \hat{\gamma} \). GLIM could be used as suggested before, or we could use a variation on the alternative iterative algorithm suggested earlier. In the latter case the algorithm is similar to that used for the doubly constrained model except that the ‘balancing step’ now only involves stabilisation of the \( \hat{\alpha}_i \) parameter estimates, whilst the likelihood maximisation step involves a more complex likelihood function and maximisation is with respect to \( \theta \) as well as \( \gamma \). We do not go into details of such parameter estimation algorithms here.

The obvious alternative to an origin constrained model is one which is destination constrained (also called attraction constrained). This is exactly analogous to the origin constrained case except it is the \( \alpha_i \) that are now modelled in terms of covariates, \( x_i^{(o)} \), measured at each origin. The model is thus:

\[ Y_{ij} = \beta_j e^{u(x_i^{(o)}; \phi)} + \gamma_{ij} + \epsilon_{ij} \]

where \( u(x_i^{(o)}; \phi) \) is some function of the vector of origin characteristics \( x_i^{(o)} \), involving a vector of parameters \( \phi \). Total predicted flows to each destination
are reproduced by the model, but not total flows from each origin. Fitting is performed in an analogous way to the origin constrained case. This model has been used in a residential location context, where there is a known distribution of employment by zone (thus the destination side is fixed) and the task is to predict the likely distribution of residences. In this case, the covariate factors at the origin might be measures of the quantity and quality of the housing stock. The model can then be used to predict the likely demand for new housing at the origin end if the pattern of job opportunities at the destination end is altered.

Finally, we have the unconstrained model where both $\alpha_i$ and $\beta_j$ are modelled as functions (usually simple) of observed characteristics at both origins and destinations. The idea is that both ‘demand’ and ‘attractiveness’ can be ‘explained’ in terms of a few measurable attributes of origin and destination locations. The model then becomes:

$$Y_{ij} = e^{u(x_i^{(o)};\phi)+v(x_j^{(d)};\theta)+\gamma d_{ij}} + \epsilon_{ij}$$

and is not constrained to reproduce the total observed flow either from any origin or to any destination. Again, fitting is by maximum likelihood using a Poisson assumption. Since no constraints are involved no alternative ‘balancing’ algorithm can be developed.

A particular form of the unconstrained model which has often been used in studies of population migration is where $x_i^{(o)}$ is taken to be simply a single variable, the logarithm of the population, $P_i$, at origin $i$, and similarly $x_j^{(d)}$ is taken to be the logarithm of the population, $P_j$, at destination $j$. An overall scaling parameter $\lambda$ is then added to reflect the general tendency for migration in the system. Hence the simple migration model that results is:

$$Y_{ij} = \lambda P_i^\phi P_j^\theta e^{\gamma d_{ij}} + \epsilon_{ij}$$

Note that in all the above variants of what we have referred to as a ‘family’ of doubly, singly or unconstrained spatial interaction models we may introduce more complex distance deterrence functions than $e^{\gamma d_{ij}}$. We may wish to include non-linear effects by, for example, including a term in $\log d_{ij}$ or effects which are different for different groups of origins and/or destinations. We will return to this point in more detail shortly, and simply note here that any of the above formulations can be rewritten by replacing the term $\gamma d_{ij}$ with some function $w(d_{ij}, \gamma)$ of a vector of distance measures $d_{ij}$ and a vector of parameters $\gamma$. No changes are required to the basic fitting algorithms, although these will clearly become more computationally intensive.

Let us now look at applying a selection of these different forms of model to our case studies. First we consider modifying our earlier doubly constrained model for the Dutch migration data, by using an origin constrained model where we model the ‘attractiveness’ of provinces for migration as a function simply of their total population. That is, we take $x_j^{(d)}$ to be a single variable, the
The analysis of spatial interaction data

logarithm of the total population (000s), \( P_j \), at destination \( j \); so our model becomes:

\[
Y_{ij} = \alpha_i P_j^\theta e^{\gamma d_{ij}} + \epsilon_{ij}
\]

The results are given in Table 9.3. Although the overall magnitude of the \( \hat{\alpha}_i \) has changed, their relative values are broadly as before. The distance parameter also remains the same. The estimated parameter for measuring ‘attractiveness’ in terms of population is highly significant, and interestingly the overall fit of this model to the observed flows is not significantly worse than that of the previous doubly constrained model (as measured by the change in ‘deviance’ discussed in Chapter 8).

**Table 9.3 Parameters of origin constrained model for Dutch migration**

<table>
<thead>
<tr>
<th>( \hat{\alpha}_i )</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Drenthe</td>
<td>0.404</td>
</tr>
<tr>
<td>Friesland</td>
<td>0.648</td>
</tr>
<tr>
<td>Gelderland</td>
<td>1.170</td>
</tr>
<tr>
<td>Groningen</td>
<td>1.105</td>
</tr>
<tr>
<td>Limburg</td>
<td>1.130</td>
</tr>
<tr>
<td>Noord-Brabant</td>
<td>1.219</td>
</tr>
<tr>
<td>Noord-Holland</td>
<td>1.763</td>
</tr>
<tr>
<td>Overijssel</td>
<td>0.725</td>
</tr>
<tr>
<td>Utrecht</td>
<td>0.703</td>
</tr>
<tr>
<td>Zeeland</td>
<td>0.470</td>
</tr>
<tr>
<td>Zuid-Holland</td>
<td>2.033</td>
</tr>
</tbody>
</table>

\[
\hat{\theta} = 0.4846 \ (0.0749) \\
\hat{\gamma} = -0.0241 \ (0.0009)
\]

The predicted flow (000s) from Drenthe to Groningen, which has a population of 532.7 (000s), is now:

\[
\hat{y}_{14} = 0.404 \times 532.7^{0.485} \times e^{-0.0241 \times 42.9} = 3.02
\]

an under prediction of the observed flow of 3.89. Under this model the total predicted flows from a province to all provinces will correspond to those observed, but the total predicted flows into a province from all provinces will not necessarily reproduce the observed numbers.

Finally, we may go one stage further and consider an unconstrained model for these data, where the origin effects are also modelled simply as a function of the total population of a province. That is, we take \( x_i^{(o)} \) to be a single variable, the logarithm of the total population (000s) and include an additional scaling parameter, \( \lambda \), to represent the overall tendency of the population to migrate.
We are then using precisely the simple migration model described earlier. This gives the results below:

\[ \hat{\lambda} = +0.0085 \]
\[ \hat{\phi} = +0.6965 \]
\[ \hat{\theta} = +0.4654 \]
\[ \hat{\gamma} = -0.0242 \]

All four parameters are highly significant. Our Drenthe to Groningen prediction, given that the population of Drenthe is 393.7 (000s), is now:

\[ \hat{y}_{14} = 0.0085 \times 393.7^{0.697} \times 532.7^{0.465} \times e^{-0.0242 \times 42.9} = 3.59 \]

a slightly better fit than before. Under this model neither total predicted flows out of, nor into, provinces will necessarily agree with those observed.

The overall fit of this model is not significantly worse than the singly constrained model given earlier (again as assessed through the change in deviance), or indeed from that of the original doubly constrained model. Since the original doubly constrained model is in a sense the ‘best we can do’ at explaining the flows (it has a separate parameter for each origin and each destination) and this is not significantly better than the unconstrained model that we have now arrived at, we can conclude that the simple unconstrained model is probably as reasonable an explanation of the migration flows as we are likely to get with these data by using gravity models. Of course, with better separation measures and more data on the characteristics of the provinces, we might well improve the level of explanation.

We may also apply an unconstrained model of a similar type to the data on London patient flows. Here there is just a single destination so that our previous model becomes

\[ Y_i = \lambda P_i^\phi e^{\gamma d_i} + \epsilon_i \]

where \( Y_i \) represents the flow to the hospital from district \( i \), which has population \( P_i \) (000s) and is at a distance \( d_i \) from it. Here we will simply use straight line distance in kilometres as the distance separation, with the distance for Hampstead (the district where the hospital is situated) being taken as half the distance of the next nearest district. Obviously, such crude distance measures will not reflect the realistic ‘cost of travel’ or ‘travel time’ in the London area, but there will be a broad relationship.

We could fit the above model to the observed flows for any particular speciality (type of treatment). However, this is equivalent to simply rewriting it as:

\[ Y_{ik} = \lambda_k P_i^\phi e^{\gamma d_i} + \epsilon_{ik} \]

where \( Y_{ik} \) is now the flow for speciality \( k \) and we have provided different model
parameters for each speciality. These parameters then characterise the ‘catchment population’ of this hospital’s workload for different specialities \( k \), in terms of a district’s population (through \( \lambda_k P_i^{\phi_k} \)) and the distance of this population from the hospital (through \( \gamma_k \)). Once fitted we can then examine for significant differences between the parameters for the different specialities and possibly simplify the model, perhaps by forming groups of specialities with similar behaviour.

We do not report the parameter estimates in detail here; however, we can summarise the general results. Firstly, there is very little evidence that the parameter \( \phi_k \) helps in this particular case. Its value is not significantly different from unity for any speciality. Hence the \( \lambda_k \) estimates obtained can be interpreted as simply ‘distance corrected’ workload rates per thousand population. The differences in these values are much as might have been expected, the highest values being in General Medicine, General Surgery and Gynaecology with lower values for more specialised treatments, such as Urology. Generally, the value of the distance parameter \( \gamma_k \) is approximately \(-0.45\) for most specialities, which, given that distances are being measured in kilometres, corresponds to a maximum catchment radius of around 9 km, before the exponential decay function becomes insignificantly small. However there are some specialities with significantly wider catchments of up to 25 km, notably specialist areas such as Neurology, Neurosurgery, Oral Surgery and Radiotherapy; the last two having the highest workload rates among the four. On the other hand, the distance parameter for Geriatrics showed a very much smaller catchment area of no more than 3–4 km.

Notice that in all the examples we have used here we have estimated a single, ‘system-wide’ distance deterrence parameter for any given type of flow. We have already alluded to the possibility of considering either origin constrained or destination constrained models which involve a distance deterrence parameter specific to each origin or each destination (or to groups of them). In other words, we may wish to generalise \( \gamma \) to \( \gamma_i \) or \( \gamma_j \). Details of such models are given in the section on further reading at the end of the chapter; we simply comment here on why they might be of interest. Consider the Dutch migration data. We have already shown how to estimate an overall distance deterrence parameter. Yet with an origin specific model we could estimate such a parameter for each province. This might reveal interesting spatial variation in distance decay effects. However, a word of caution is needed here. When such origin specific models are fitted in practice it is invariably found that the distance deterrence effect is greater in peripheral regions and lower in more central areas. For example, we might obtain a \( \gamma \) of \(-0.01\) for those leaving Utrecht, in the centre of Holland, and a value of \(-0.03\) for those from Limburg in the south, were we to fit such a model. The interpretation of this might be that people leaving Limburg were more constrained, or deterred, by distance, than those from Utrecht. This seems implausible. What we are observing here is a misspecification of the spatial interaction problem. It has led writers such as Stewart Fotheringham to propose another class of model, those known as competing destination models.
In such models, we imagine that destinations 'compete' with each other for interactions, and we build into our spatial interaction model some variable to measure this. The interpretation of such competition depends upon the context. In retailing we might find that an isolated centre captures local trade because of the lack of competition; in migration studies we might imagine that people first select a region (for example, the south of Holland) and then, within that broad region, a particular destination. If, within the broad region, there are relatively few alternatives, we can expect such destinations to draw more migrants than those in regions where there are many possible destinations. As a consequence, we can build into the modelling framework a variable to capture this destination competition. One suggestion is a measure of relative accessibility of a destination, such as:

\[ q_j = \sum_k \frac{w_k}{d_{kj}} \]

where \( q_j \) is the accessibility of \( j \) to all other destinations and \( w_k \) is the attractiveness of the \( k \)th competing destination.

Although we have made reference already to the measurement of spatial separation used in spatial interaction models it is worth commenting a little more fully on this before we leave the subject. Just as we can envisage a series of factors contributing to the attractiveness of a destination, or the propensity of an origin to transmit flows, so too we can envisage a variety of factors that go to make up some suitable measure of spatial separation. In the transport planning literature such a measure is known as generalised cost. This could embrace travel time, \( t_{ij} \), factors that are proportional to actual distance, \( d_{ij} \), and the costs, \( w_{ij} \), of parking, waiting, and so on that all contribute to the cost of overcoming distance. These elements can be combined to give:

\[ c_{ij} = \theta_1 t_{ij} + \theta_2 d_{ij} + \theta_3 w_{ij} \]

where \( c_{ij} \) is generalised cost. We would then simply use \( c_{ij} \) instead of \( d_{ij} \) in all our previous models. The parameter \( \theta_2 \) translates distance into money (for example, the costs of travel by bus, per kilometre), while the other parameters, \( \theta_1, \theta_3 \), represent the valuation of travel and waiting time. We shall return later to the issue of using distances and travel times along networks as measures of spatial separation. But what the generalised cost does make clear is that \( c_{ij} \) is not necessarily the same as \( c_{ji} \). When simple Euclidean distances are used in spatial interaction modelling (as they often are) then spatial separation is, by definition, symmetric.

Another problem in using any measure of spatial separation that arises in most practical applications concerns the definition of \( d_{ii} \) or \( c_{ii} \): how are we to measure this where we are dealing with zones and there are intra-zonal flows, \( y_{ii} \)? In the models we have fitted to the Dutch migration data for illustrative purposes, we have simply taken the intra-zonal distance to be zero (note that this does not imply the predicted flow is zero). However, in general this may
not be acceptable, since it does not reflect the fact that there may be more deterrence to intra-zonal flow in one area than in another. There are no hard and fast rules here, though some theoretical results are available for zones that are circular or rectangular. If the zone is approximately circular, one suggestion is to take two thirds of its 'radius' as an intra-zonal distance.

Before we leave this section on the modelling of spatial interaction data, we should also comment on one general feature of the methods that departs markedly from the approach which we have presented earlier for the analysis of other types of spatial data. In previous chapters, whether dealing with point patterns, spatially continuous or area data, we were careful to maintain a balance in our models between both first order and possible second order effects. Indeed, on occasions we have gone to great lengths in some models to ensure that we have adequately allowed for second order effects. That is, we have modelled a spatial covariance structure in residuals about the mean, as well as modelling the mean or first order properties of the process. In this chapter we have focused entirely on modelling first order effects in spatial interaction data. Our family of gravity models are models for $\mu_{ij}$, the mean flow from $i$ to $j$. We have not commented at all on the possibility of spatial covariance in fluctuations about this mean. In fact, the maximum likelihood methods for estimating the parameters of gravity models make the explicit assumption that fluctuations about the mean are independent; that is, no second order effects are present.

Given our comments on various kinds of spatial models throughout this book, the reader should be somewhat reluctant to accept the proposition that spatial interaction data will have no spatial covariance structure. At the very least, we should explore and analyse the estimated residuals, $\hat{\epsilon}_{ij}$, for possible spatial dependence. In other words, for each predicted flow there is an associated error term, the difference between the observed and the modelled flow. The assumption of no second order effects means that the residual $\epsilon_{ij}$ is, for example, uncorrelated with $\epsilon_{ik}$. But it may be the case that if we are over predicting the flow from $i$ to $j$ we are also over predicting that from $i$ to $k$. As yet, however, there has been little work done on this problem of spatially correlated errors in spatial interaction modelling. Some of the references at the end of this chapter discuss adjusting the Poisson assumption to allow for such effects as well as 'over dispersion' of flows (larger variance than expected under a Poisson assumption). Such modifications increase the difficulties of parameter estimation, and techniques such as pseudo likelihood are then required to address such problems.

9.5 Related methods in the analysis of spatial interaction

So far, we have considered a variety of spatial interaction models and have illustrated their uses. In this section we want to consider some further issues related to the subject of spatial interaction. These take us beyond the description and modelling of observed flows. We have alluded already to the
possible use of singly constrained gravity models in determining the optimal location for new retail development. In this section we shall take that discussion further and consider in more detail the general problem of optimally locating facilities and investigating their catchment areas. Such problems are known as location-allocation problems.

Location-allocation modelling is very often concerned with locating facilities on a network rather than in continuous space. We have already referred to networks on a number of occasions. For example, we noted early in the chapter that spatial interaction modelling is often carried out in situations where origins and destinations are nodes on a transport network and flows are restricted to the arcs of that network. Networks provide a rich source of analytical problems in their own right, many of which are directly related to issues which arise in spatial interaction or location-allocation modelling. Some are concerned with finding a pattern of flows through a network that is in some sense optimal among all sets of flows which would satisfy some overall requirement. For example, we might wish to minimise the total cost of transporting goods between locations, given some requirement to provide defined levels of supply at particular nodes in the network. As we shall see, there turns out to be some relationship between such ‘classical’ transportation problems and the entropy maximisation formulation we have already considered in developing the general mathematical form of the gravity model. Other classes of network problems to which we refer are concerned with finding shortest paths, optimal routes, or maximal possible flows through a network.

Although location-allocation, transportation and other network problems are still very much of a spatial analytical nature, we shall see that they take us more into the realms of mathematical optimisation techniques and away from the area of statistical description and modelling of observed spatial data which has been the subject to which we have so far confined ourselves throughout this book. Many of the problems we will discuss, both within a location-allocation and a network context, can be formulated as requiring the optimisation of some objective function, subject to a set of constraints. In general such problems are known as mathematical programming problems. Where the objective function and constraints are all linear functions of the variables with respect to which optimisation is required, the problems are known as linear programming problems. In cases where such variables are restricted to have only integer (whole number) values, as they will be in several of the situations we will discuss, the problems are referred to as integer programming problems.

Mathematical programming in general, and linear and integer programming in particular, are extensively researched areas in their own right. Much of the subject of Operations Research (OR), is devoted to the study of such problems and very efficient algorithms have been developed for their solution. However, they rank amongst the most computationally intensive mathematical problems that exist, and exact solutions to many of the problems that we formulate in this section are not achievable with current methodology for anything other than relatively small networks. As a result, such problems are often solved
using what are called heuristics, sets of rules that give a ‘good’, but not necessarily optimal, solution to the problem.

Given the substantial nature of this area, combined with the fact that it is somewhat tangential to our main theme throughout the book, we do not deal with location-allocation modelling and network optimisation methods in any detail; indeed, we would require another book to do such methods justice. We simply review and formulate some of the most basic models and point out the links between some of these and the modelling of observed spatial interaction data, considered earlier. We provide references at the end of the chapter for the reader who wishes to follow up the techniques in more detail.

9.5.1 Location-allocation problems

We suggested earlier that an origin constrained spatial interaction model could be used to make predictions, \( \sum_i \hat{y}_{ij} \), of the volume of flow arriving at destination \( j \) given a pattern of demand at the origins and an attractiveness function for each of the destinations, which might typically include covariates such as retail floor space, parking facilities and so on. In a retail context such predictions allow us to derive estimates of likely sales at centre \( j \) under various different scenarios. Such ‘shopping’ models can then be used in a forecasting context. For example, if we add to the volume of retail floor space at \( j \), a new set of predicted flows is then generated, along with new predictions of the distribution of retail sales and the revenue that suppliers at \( j \) can expect to achieve. Alternatively, we can model the changes to the flow of retail expenditure if the demands at the origins are modified, or if the transport network (i.e. distance or cost of travel) changes. The spatial interaction model represents the ‘system’ and allows us to predict how small changes to one component of the system (such as one origin demand, one change at the destination end, one modification to the transport network) will feed through to possibly have indirect effects at all locations in the system.

This understanding that we might use certain kinds of spatial interaction model to predict ‘what might happen if’, immediately raises the possibility of using them to help in optimising the placement of retail outlets. Suppose we wish to locate a new retail outlet in an urban area. An origin constrained model can be fitted to flows of customers using both our existing facilities and those of our major competitors. Then, given future patterns of demand at origins perhaps based on estimated expenditure on retail goods by residents in zone \( i \), we could predict likely expenditure on retail goods generated at any potential location in the urban area, simply by inserting the set of \( d_{ij} \) that represent the cost of residents travelling from homes in zone \( i \) to the hypothetical destination \( j \). The predicted total flows of expenditure at all such hypothetical locations can then be mapped as a spatially continuous revenue surface (though one that currently ignores the costs of locating at \( j \)); the peak on this surface is then the optimal location, in the sense of maximising revenue.

Many ‘classical’ location-allocation models, such as those based on central
place theory, tend to assume that all individuals who 'flow' (e.g. shoppers) will always visit the closest centre or facility that provides the service they seek. In other words, given a particular placement of facilities, the assignment or 'allocation' of individuals to facilities is then deterministic, or entirely predictable. As a result, the catchment area of a facility is unambiguous and well defined. The kind of spatial interaction models we have been discussing throughout this chapter offer a much more natural, probabilistic, assignment of demand to destinations; a better way of characterising the catchment area of a facility. They recognise that while some shoppers in a zone will visit one centre, others will go elsewhere. Distance, along with other factors, constrains, but does not necessarily determine, the pattern of interaction. For these kinds of reasons a current area of research concerns how to bring together 'classical' location-allocation modelling and spatial interaction modelling, a point we will return to later.

First, however, we need to review briefly the more 'classical' models for optimal facility location. In introducing the subject we considered one possible criterion for determining optimal location of a facility such as a retail centre, that of maximising total possible flows or total possible revenue associated with the facility. There are several other criteria for location that we could consider, and location-allocation modelling in general is concerned with the optimisation of various such criteria, subject to assumptions concerning the factors that determine the 'catchment population' of any potential facility location and constraints concerning demand for the service provided at the various 'origins' involved. As mentioned above, many of the techniques focus on constrained optimisation methods for particular criteria, under deterministic assumptions about catchment populations, usually that individuals will choose the 'nearest' facility. We have already made the point that the 'lessons' of spatial interaction models suggest that the behaviour of flows is potentially more complex and probabilistic. Nevertheless, one might argue that this is essentially a question of how one defines catchment population and does not ultimately alter or detract from the overall location optimisation problem to which much of the work in this area has been devoted.

To gain some insight into location-allocation models it is useful to bear in mind the basic structure of the problem they address. In all cases there is demand for some good or service and this demand is allocated, or flows, to particular centres according to a pre-defined set of rules which determine 'who will flow where'. We then wish to locate centres or facilities in such a way that some criterion is optimised, subject to the given demand. Very often there is only a certain discrete set of potential sites at which to locate the facilities and the assumption is made that demand can also be taken to occur at point locations. In other words, the location problem is conceived as being one of placing facilities at certain nodes on a network. This is certainly the case in most of the traditional models, although work has also been done on locating facilities in continuous space.

Reference back to two of our case studies is perhaps instructive here in order to illustrate the problem addressed. In the bank account example, which is very
similar to the general retail situation, we might want to locate a new branch that captures as much account business as it can. In the case of patient flow, where we know, or can estimate, the demand for health care, we might wish to determine an optimal location for a new hospital. The reader may view this latter application as somewhat unrealistic in the context of the developed world, where rationalisation (closure!) of hospital sites seems to be the order of the day, but nevertheless, it is of obvious importance in the developing world, where scarce resources must be targeted as carefully as possible when devising optimal location for health care facilities.

In seeking an optimal location we need an objective function. In the retail case this might simply be to maximise profits. But in general we can try to optimise (minimise or maximise) other objective functions. This is particularly the case in the context of locating public facilities, where we may wish to locate public libraries, health clinics, fire stations, or recreation centres to best serve the population. One widely used objective function in these contexts is to assume that individuals will always travel to their nearest facility, and then to attempt to choose the location of facilities by minimising the total distance travelled (travel cost) in the system (equivalently, we could think of this as a problem of maximising system-wide accessibility). In this case we may formulate the problem of optimally locating \( p \) facilities in a system with \( n \) possible nodes as:

\[
\min \left\{ \sum_{i=1}^{n} \sum_{j=1}^{n} a_i I_{ij} d_{ij} \right\}
\]

subject to the constraints:

\[
\sum_{j=1}^{n} I_{ij} = 1 \quad i = 1, \ldots, n
\]

\[
\sum_{j=1}^{n} I_{jj} = p
\]

\[
I_{ij} \leq I_{jj} \quad i \neq j
\]

Here the given demand at \( i \) is represented by \( a_i \) and as usual, \( d_{ij} \) is some measure of the cost of reaching \( j \) from \( i \). \( I_{ij} \) is a set of indicator variables in respect of which the minimisation is performed, where these take the value 1 if all demand from \( i \) is allocated to facility \( j \) and 0 otherwise. The second constraint ensures that exactly \( p \) facilities are located and the third that no demand is met from \( j \) if there is no facility located there. The first set of constraints ensures that \( a_i \), the demand from \( i \), is met in full, since it ensures that demand from each \( i \) is allocated to exactly one \( j \). The precise nature of the given demand will depend upon the particular problem. If we are locating fire stations it might relate to the number of domestic and non-domestic properties in a zone; if we are locating public libraries it might simply be related to total population.
When the above model is used to determine the location of a single centre (the case where \( p = 1 \)) it is essentially identical to the so-called Weber problem in industrial location. More generally, however, we shall wish to use it to locate \( p \) centres, the so-called \( p \)-median problem. It seeks the location of \( p \) supply centres to minimise the total cost of travel whilst satisfying demand. Solutions to this problem have been developed both in a continuous setting (where we assume that centres are free to locate anywhere on a plane) and a discrete context, where both centres of demand and the set of feasible locations are points on a transport network (the case formulated above). In general, exact solutions to the latter problem are computationally infeasible for problems of a realistic size. However, good heuristics are available to find acceptable approximate solutions.

As an example of the use of these methods, consider the problem of locating \( p = 6 \) health centres in the district of Salcette Taluka, in the west of Goa, India. The current health care delivery system already comprises six centres; are these optimally located, according to the objective function defined above? The solution to the \( p \)-median problem suggests that two health centres, including that in Margao, the major urban centre in Goa, are optimally located, but that four are not. As can be seen from the mapped outcome (Figure 9.6) the \( p \)-median solution suggests that the east of the study area is rather poorly served. Note that the \( p \)-median solution also provides the allocation of demand (population) to the optimally located health centres. However, the solution shown here used simple straight-line distance as a measure of spatial separation and took no account of the capacity of centres to deal with patients.

The \( p \)-median problem has been widely used in a variety of planning contexts, not simply in health care planning. There are a number of variants on the basic problem. For example, it may be adjusted to take account of varying installation costs of facilities at different nodes, or capacity constraints on the facilities that are located. It can also be extended to include constraints on the maximum distance that individuals will travel to facilities. This last extension raises an important point concerning the \( p \)-median formulation. Whilst it generates an 'efficient' solution (in the sense of minimising aggregate travel costs) it does not necessarily generate an 'equitable' solution. In particular, if we are dealing with a situation where demand is inelastic, that is, where clients require the service regardless of the cost of getting it—health care or fire cover, for example—we find that the \( p \)-median solution tends to ignore the needs of those located furthest away from the chosen optimal location(s).

In order to reduce the variability in the accessibility of origin zones and individuals living therein, alternative objective functions have been suggested. One possibility is to choose a set of centres which minimise the maximum distance separating any user from the nearest facility. This is called a minimax solution. Another solution is via what are called covering objectives. With these, a standard of service is defined which must be met for all origin zones in terms of a distance (or time) limit from facilities. So, a set of locations is sought such that the entire demand is 'covered', or met, within a critical distance or time. A special and commonly used example of this, the maximal covering problem,
Fig. 9.6 Optimal location of health centres in West Goa. Reprinted from Hodgson, M.J. (1988) Social Science and Medicine 26 153–61, with kind permission from Elsevier Science Ltd, The Boulevard, Langford Lane, Kidlington OX5 1GB, UK.

seeks to find the locations of \( p \) facilities so that the number of people within some specified critical distance or time threshold, \( h \), is maximised. Mathematically we may formulate this problem as:

\[
\min \left\{ \sum_{i=1}^{n} I_i(h) a_i \right\}
\]

subject to the constraints:

\[
\sum_{j:d_{ij} \leq h} (I_j + I_i(h)) \geq 1 \quad i = 1, \ldots, n
\]

\[
\sum_{j=1}^{n} I_j = p
\]

As before, the given demand at \( i \) is represented by \( a_i \) and \( d_{ij} \) is some measure of the cost of reaching \( j \) from \( i \). However, we now have two sets of indicator
variables. \( I_i(h) \) is 1 if node \( i \) is not within \( h \) of any facility located (i.e. it is not 'covered'). \( I_j \) is the set of indicator variables in respect of which optimisation is carried out, taking the value 1 if a facility is located at \( j \) and 0 otherwise. The objective function is expressed in terms of minimising the amount of demand not covered. The first set of constraints determines whether demand points are covered or not whilst the second ensures that exactly \( p \) facilities are located.

The parameter \( h \) takes on an interesting role here. Where \( h \) is relatively large, the solution to the problem becomes similar to that of the minimax solution; as \( h \) shrinks, however, the solution becomes less equitable, since the optimal location is drawn towards the zone with the greatest demand; the demands of other zones are ignored. One of the attractive features of the maximal covering problem is that it can be used to explore just how much cover a certain value of \( p \) can provide. Repeated solution can therefore give insight into just how many facilities are required to provide some defined level of ‘cover’. This latter problem can also be formulated directly as a modification of the basic problem if required—the so-called location set covering problem. There are several other variants of the basic maximal covering problem. As with the \( p \)-median problem, exact solutions to the maximal covering problem and its variants are computationally infeasible for networks of even moderate size. Heuristics are available to find acceptable approximate solutions.

How do these sorts of ‘classical’ mathematical optimisation problems relate to the spatial interaction models we have described earlier in this chapter? We have already touched on this earlier. The optimising criterion in the classical OR models assumes that people visit the closest centre; in other words, that the allocation of demand for any particular facility locations is deterministic. But we have already seen that spatial interaction models generate a probabilistic allocation of demand to centres of supply. Because of this, current research in location-allocation modelling attempts to link these models with spatial interaction models that capture the more complex behaviour of trip-makers. In particular, location-allocation models have been devised that both locate \( p \) centres \textit{and} are consistent with probabilistic travel behaviour. One such model sets up an objective function which essentially involves two terms. The first represents average travel costs and is similar to the objective function in a \( p \)-median problem. The second involves a distance deterrence parameter and may be interpreted as a measure of the costs or ‘disutility’ of having a set of possible centres dispersed over space. Together, both these terms provide a measure of overall consumer costs. We do not give details of the explicit formulation here.

In what other ways can we extend location-allocation models, other than by trying to make them more consistent with spatial interaction models? First, we should recognise that, in some cases, we shall want to place constraints on the ability of centres to cope with demand; in other words, we might want to attach \textit{capacity constraints} to centres. These might represent numbers of available bed spaces in hospitals, for example. The ‘classical’ location-allocation models, such as the \( p \)-median problem discussed earlier, generate an ‘un capacitated’ solution. However, more realistically we might also wish to incorporate a capacity constraint. This is generally straightforward both for the \( p \)-median
and for other models we have discussed, although clearly such additions make solutions computationally harder to obtain.

Another possible extension is to develop *hierarchical* location-allocation models. With these, we want to locate not just a single set of facilities, but facilities that occupy different levels of a hierarchy. To return to the problem of locating health centres in part of Goa, for example, one set of health clinics might offer basic nursing and primary care while at a higher level of the hierarchy we might have one or more health centres providing more staff and a fuller range of health care services. It is usual to assume that all low level functions are also catered for in the higher level clinics. Now, in the $p$-median problem, as we have already seen, people are assumed to make their way to the nearest centre. But this may be rather unrealistic. Might not patients prefer to travel to a larger hospital, one with more staff, a wider range of services and a facility located in a larger urban area, allowing for the combining of a visit to the health facility with other purposes? Location-allocation models have been formulated to deal with this kind of situation. They consider optimal siting of fixed numbers of facilities at each level of a hierarchy of different types. We do not give details of such models here, but simply note that most of the models that we have previously mentioned can be extended to a hierarchical context.

Before leaving location-allocation models we should draw attention to the fact that often in facility location we may wish to optimise not just a single objective function but, rather, a set of such functions. This gives rise to what is known as *multi-criteria* problem solving. Here, a decision maker wishes to determine a configuration of facilities that tries to optimise not only, say, aggregate distance travelled (the $p$-median problem) but simultaneously also various equity criteria. The extent to which such multi-criteria situations can be addressed by mathematical optimisation techniques is somewhat limited. Certainly it is perfectly possible to set up multi-criteria objective functions, but the real problem is in reliably quantifying and making explicit a decision maker’s attitude to competing objectives. To what extent are they prepared to trade off one criterion for another? Are such trade-offs constant over the whole range of possibilities or are there special cases where they would not apply? Quantitative methods do exist to address such questions, but they carry with them a set of behavioural assumptions which some might find hard to accept.

This brings us to a final and related general point. We should emphasise that although location-allocation modelling by its nature focuses on the optimisation of pre-defined criteria, the value of many of these models is rarely the optimal solution they provide, but rather their ability to explore and compare solutions. Facility location in both the public and commercial sectors is a complex decision involving many interacting factors. Optimisation models can provide useful input to that decision making process, but alone they will never provide an ultimate answer. Where they are most useful is in guiding decision makers towards ‘good’ solutions, allowing the implications of potential solutions to be objectively evaluated and in exploring how sensitive various solutions may be to assumptions or estimates incorporated into the analysis.
9.5.2 Network problems

In the previous section we have reviewed methods for finding an optimal set of locations, where 'optimal' related to various specified criteria. In this section we shall consider the locations of facilities in a network to be given, and consider instead a different class of problem concerned with finding various kinds of 'optimal' paths through the network connecting these facility locations. For example, we might seek one path through the network that minimises the cost of travel between two locations. Alternatively, we might consider the problem of optimally choosing an entire configuration of flows in order to move people or goods between a set of supply nodes in the network and a set of demand nodes in the most 'efficient' way—the so-called 'transportation problem', mentioned earlier. We have already referred to this general class of problem as 'network problems'.

When we considered spatial interaction models earlier, our objective was to obtain sensible and succinct mathematical descriptions of observed spatial interaction. In this section, as in the last, we are not involved with description of a pattern of flows, but rather with optimally assigning a pattern of flows. However, we shall see that there are relationships between network problems and spatial interaction models. One relationship that is immediately apparent is that the analysis of various kinds of paths in a network, such as finding those associated with shortest distance or minimum cost of travel, will have a direct bearing on the sort of spatial separation or distance measures that might be used in spatial interaction models pertaining to that network. However, we shall also see that there are some less obvious relationships, for example between transportation problems and spatial interaction models.

Before considering issues concerned with paths and flows through a network, we first make a few remarks about network design. Here, we are given a set of locations to be connected with some transport or communications network. How are we to do this? Again, we need the concept of an objective function. One possibility would be to seek a maximally connected network, with each place connected directly to every other; clearly, this would minimise users' transport costs. Conversely, we could design a network that minimised construction costs, a network that connected all places but with the minimum possible length (or cost) of route; the solution here creates a minimum spanning tree, which we first encountered in a different context in Section 6.3.5. We refer the reader to the description there of how it is obtained. Of course, in reality, transport networks are invariably designed as a compromise between these various objectives.

Suppose now that the network is given and that each link or arc on this network has a value (distance, cost, or time) associated with it (sometimes this is referred to as the impedance on an arc, since it impedes the flow). A whole variety of different optimisation problems may now be posed. For example, at the simplest level, we could ask which is the path or route between a given pair of places, or nodes, on the network that minimises total travel cost (or distance, or time)? Such problems are referred to as shortest path problems and one
commonly used algorithm for solving such problems is known as Dijkstra's algorithm which maps out a minimum cost route (or shortest path if measured as total arc length) from the origin to the destination. We do not give details of the algorithm here; they may be found in any elementary OR text. Another problem involves finding the maximum flow that is possible from an origin to a destination, given capacity constraints attached to each arc or link. This problem can be combined with the shortest path problem to ask: what is the least cost transportation arrangement from i to j, given a number of items to transport and a set of capacity constraints attached to each arc? That is, given N items (people or goods, for example) that need to go from i to j, what is the path through the network such that total transport costs (distances) are minimised without exceeding the capacity constraints on any link? This least-cost, maximum-flow problem may be solved using the so-called out-of-kilter algorithm which partitions the total flow N that must go from i to j among different possible paths in the appropriate optimal way. Methods such as these are of enormous value in many areas of application; an obvious one being emergency evacuation, where people must be evacuated from a hazard incident zone to one or more reception centres. This must be done in an efficient way, minimising time costs for example, but it will also need to allow for the fact that arcs will be capacitated; that is, limited in terms of the numbers of vehicles with which they can deal.

A further class of network problem involves the derivation of tours in the network with various desirable properties. A tour, as its name suggests, is a route from an origin i that passes through all other nodes before returning to i; for less obvious reasons, it is also called a Hamiltonian circuit. If we want to find the shortest possible such circuit in a given network, we are involved in what is classically referred to as the travelling salesman problem. This can be formulated as an integer programming problem similar to some of those we encountered in the previous section on location-allocation modelling. However, formulation of a problem is not the same as solving it! Exact solutions to this combinatorial optimisation problem are computationally infeasible for all but the smallest of networks, even on the fastest computers. As a result, such problems are solved, like many of our previous location-allocation problems, by using heuristics, sets of rules that give a ‘good’, but not necessarily optimal, solution to the problem. For example, one possible heuristic for the travelling salesman problem is at each stage to visit the nearest unvisited place. This is not a particularly good heuristic, since it results in longer trips in the later stage of a tour; better heuristics can be devised as a modification of this basic idea.

As we have seen, many of the optimisation problems we have considered, both within a location-allocation and a network context, reduce to the constrained optimisation of some objective function. We commented earlier that such problems are known, in general, as mathematical programming problems and come in various varieties such as linear programming and integer programming. The latter refers to the case where variables in respect of which optimisation is performed can only take integer (whole number) values, which
has in fact been the case in several of the situations we have discussed. One final class of mathematical programming problems we wish to mention is the so-called *transportation problem*. The structure of this turns out to be not so very different to the spatial interaction problem we began with in this chapter, although the objective is rather different. Recall that, in developing the general spatial interaction model earlier, we imposed a total cost constraint; we wanted to derive the most likely pattern of flows whilst preserving the observed total cost of travel in the system as well as the observed total flow from each origin and to each destination. Suppose now we set up the following constrained optimisation problem:

$$\min \left\{ \sum_{i=1}^{m} \sum_{j=1}^{n} y_{ij} d_{ij} \right\}$$

subject to:

$$\sum_{j=1}^{n} y_{ij} = a_i \quad i = 1, \ldots, m$$

$$\sum_{i=1}^{m} y_{ij} = b_j \quad j = 1, \ldots, n$$

The two constraints are identical to the origin and destination constraints that appeared in deriving the general spatial interaction model. However, instead of wishing to derive a most probable pattern of spatial interaction assuming a fixed total cost of travel and given total numbers that must flow from each origin and to each destination, we now wish to find an optimal pattern of flow (one with minimum total cost) subject to the origin and destination constraints. We know the locations of the origins and destinations (so we are not dealing with a location problem); we want an optimal allocation of flows, optimality here meaning minimisation of total transport cost.

This transportation problem arises in all sorts of situations, involving both private and public sector contexts. For example, we might take a set of regions producing known quantities of gas, together with a set of demand zones with known requirements. Or, we might have a set of residential zones, supplying known numbers of students to schools ‘demanding’ pupils. A solution to the transportation problem will, in both cases, yield a configuration of flows from origin to destination that minimises total transport cost. In the school districting example, it will result in a set of catchment areas for those schools. In some situations, interest in solutions to the transportation problem centres on its use in a policy context to redesign flow patterns; in other cases, however, it may be used simply as a benchmark against which to measure how far actual flows depart from that of system-wide efficiency. Again, we do not give details here concerning the solution to such problems; suffice it to say that most transportation problems are now solved using general purpose packages developed for mathematical programming and, more especially, linear programming.
Before leaving this brief consideration of network problems in general and of the transportation problem in particular, we comment again on the similarities between the structure of the latter and the one we came across in deriving the general spatial interaction model. In one sense we have come full circle in the chapter—the system-wide cost of travel which constitutes the objective function in the transportation problem was one of the characteristics of an observed pattern of flows that we chose to preserve in deriving a general mathematical form for a spatial interaction model. We also chose to preserve total flows to destinations and from origins, the two sets of constraints which also appear in the transportation problem. What, then, is the relationship between these two different problems, the one concerned with a sensible model to describe an observed pattern of flows subject to preserving observed origin and destination total flows, the other with deriving an optimal set of flows subject to the same constraints? The answer lies in a reciprocal relationship between the distance deterrence parameter in the spatial interaction model and the total observed cost of travel in a given system of flows. If we observe a set of flows whose total cost of travel far exceeds the least possible cost as indicated by the solution to the transportation problem with the same origin and destination constraints, then it must be because cost of travel is not at a premium for individuals flowing in the observed system. In other words, we will get an estimated distance deterrence effect for that system which is quite low. On the other hand, if the observed flows have a total cost of travel near to the theoretical minimum represented by the transportation solution, then the distance deterrence effect must be high in the system of observed flows. In this sense the optimal transportation solution for given origin and destination total flows represents an upper bound for the size of the distance deterrence effect in an observed system with those same origin and destination totals. This makes intuitive sense; as distance becomes more and more of a deterrent towards interaction, the observed pattern of flows should become increasingly similar to that which minimises total travel costs.

9.5.3 Software environments for spatial interaction modelling

Given that the focus of our book is on interactive spatial analysis we should say something about the kinds of software environments that are currently available for solving spatial interaction, location-allocation, and network problems. As with the situation described in Chapter 2 in respect of more general spatial data analysis, until quite recently software for solving spatial interaction, location-allocation and network problems tended to rely on the running of non-interactive computer programs. That is, data files were prepared (for example, of origin and destination factors and distance matrices), and programs run to generate print-outs that contained little, if any, graphic display. Of course, such software was valuable; it allowed both academics and planners to generate solutions to quite complex problems, and to experiment with different scenarios by submitting another program to the operating system.
However, in the last few years software environments have changed dramatically, and, partly due to the advent of GIS, a number of packages and modules have appeared to greatly enhance the usability of the methods we have described earlier in this chapter. We comment briefly on some examples, recognising that these are merely a few illustrations. Other software is being developed, or will surely develop rapidly, utterly transforming our ability to perform sophisticated analyses. We focus on software developments primarily directed towards spatial applications and leave aside the considerable and continuing advances in general mathematical optimisation software, which also has a significant bearing on our ability to tackle spatial interaction problems.

The GIS software package IDRISI, to which we referred in Chapter 2, offers built-in modules for performing multi-criteria decision making and for solving least-cost routing problems (for example, by attaching weights to different classes of land use so that the route taken by a new transmission line avoids highly weighted areas of prime land). However, the versatility of IDRISI is that it permits users to link in other software. For instance, some researchers have attached simple spatial interaction and location-allocation models to IDRISI. One module determines the accessibility of facilities to a set of demand locations, given a fixed form of distance deterrence and a pre-specified $\gamma$ parameter. Other modules solve the p-median problem and the maximal covering problem. The solutions to these problems may then be displayed using the graphical capabilities of IDRISI.

Another comprehensive package for spatial interaction modelling and network analysis is called TransCAD. This has facilities for fitting all the spatial interaction models we have outlined, together with various modules for network analysis and location problems. Given its use of detailed digital road network databases it offers considerable functionality for the sorts of modelling we have emphasised in this chapter.

At the time of writing, ARC/INFO, one of the most widely used proprietary GIS, does not offer the ability to solve location problems. It does, however, give functions for solving shortest path and allocation problems, in discrete space. The network modules allow, among other things, functions for attaching impedances to arcs (such as estimates of travel time, and flow capacities), barriers at junctions to prevent movement, and constraints at junctions to reflect restrictions on certain turns at those junctions. Shortest path algorithms may then be invoked to find least-cost routes through the network. Allocation problems are solved by first fixing the locations of facilities at nodes on the network and also attaching to those facilities a capacity on the volume of activity each can handle if this is required. Arcs are then allocated to the nearest facility along the least-cost route, until the maximum impedance limit is reached. For example, if we wished to find all arcs on a road network that were within 30 minutes (the maximum impedance) of a major hospital, the system would find all arcs that could be reached from a hospital within that time. Alternatively, given a set of schools each with a fixed capacity we could allocate ‘arc demand’ (the number of students living along
particular streets) to those schools until the school capacity is met or the maximum impedance (distance or travel time) is met, whichever occurs first.

One of the main advantages of performing this kind of analysis within a GIS environment is that a good system will allow experimentation: the posing of "what if" questions. What if we convert a set of routes into one-way systems; or if we remove a particular school or hospital at a node on the network; or if we increase the capacity of a particular facility? The interactive nature of modern software means that answers to such questions can be obtained quite rapidly. Given a suitable graphics environment, the analyst can select particular arcs, nodes, or centres for modification and then 'see' the results of such alterations to the database. We emphasise that this is only possible with good, interactive software, as well, of course, as the availability of good digital data, most obviously on the transport network. As we pointed out in Chapter 2, such data are becoming widely available, particularly in the developed world.

The ability to blend good digital data with imaginative software lies at the heart of GIS. But the link to spatial interaction, location-allocation, and network modelling, provides a much more powerful analytical environment. Indeed, these sorts of links give rise to what has become known as a \textit{Spatial Decision Support System} (SDSS). Here, data can be readily interrogated and updated, but the system allows for forecasting, impact analysis and optimisation. Coupled to software that allows for genuine interaction between users and their data, such SDSS may indeed prove to be powerful tools.

\section{9.6 Summary}

In the early sections of this chapter we were concerned again with the main theme of our book, that of the statistical description and modelling of observed spatial data, considering in this case data on spatial interaction. We followed the structure established in earlier chapters, talking briefly about methods for visualising and exploring such data before moving on to more formal statistical modelling techniques. We introduced the general spatial interaction or gravity model and discussed methods for fitting such models to observed flows. We considered variants of the basic doubly constrained model and the various potential applications of such models.

In the final sections of the chapter we broadened this discussion to consider various optimisation problems related to the analysis of spatial interaction; problems such as the optimal siting of facilities, optimal paths and tours in networks and the 'classical' transportation problem. In doing so, we moved away from the general theme of the book, which was concerned with spatial data analysis and more into the area of spatial analysis in general. Statistical methods moved into the background and the techniques of Operations Research became important; data modelling was replaced with constrained optimisation. This probably means it is time to stop writing!
9.7 Further reading

The Swedish airline data come from:

The data on migration between Dutch provinces are taken from the following source. As mentioned in the text, the data we include are total flows; but the publication gives data disaggregated by age, so interested readers might want to create further data sets relating to retirement migration, or movement of those aged under 30, for example.


The data on hospital flows are taken from the so-called Körner Minimum Data Set collected for all hospitals in England and Wales, combined with Hospital Activity Analysis returns, for the Regional Health Authorities embracing the London region. The data are of historical interest only, relating to the mid-1980s.

The data on current accounts are described, and models fitted, in:


On the visualisation and exploration of spatial interaction data, see:


Good introductory material on spatial interaction models is to be found in the following:


Other, more advanced, references on spatial interaction modelling include:


A good reference to general linear modelling, including use of the statistical package GLIM is:

General references on parameter estimation in gravity models include:


Excellent introductions to location-allocation models are given in:


The application of the p-median problem to health care delivery in Goa, together with an outline of the hierarchical model, comes from:


For a detailed mathematical development of location-allocation models see:


Good general references that cover all or some of the techniques for mathematical optimisation, network problems, and specifically the transportation problem, are:


In terms of software we have already given some references to GIS products in Chapter 2. An early set of computer routines for solving a variety of location-allocation problems is provided in:

**Goodchild, M.F.** and **Noronha, V.T.** (1983) *Location-Allocation For Small Computers*, Department of Geography, University of Iowa, Monograph No. 8, Iowa City, Iowa.

The addition of modules to IDRISI for solving location-allocation problems is described in:


And, for a statement of the importance and nature of Spatial Decision Support Systems, see: