International Institute for
Applied Systems Analysis
IIASA www.iiasa.ac.at

Structurally Stable Transport Flows and Patterns of Location

Puu, T.

IIASA Research Report
November 1982

STRUCTURALLY STABLE TRANSPORT FLOWS AND PATTERNS OF LOCATION

Tönu Puu
Department of Economics, Umeå University, S-901 87 Umeå, Sweden

RR-82-42
December 1982

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
Laxenburg, Austria

International Standard Book Number 3-7045-0047-X

Research Reports, which record research conducted at IIASA, are independently reviewed before publication. However, the views and opinions they express are not necessarily those of the Institute or the National Member Organizations that support it.

FOREWORD

The work on regional development at IIASA is devoted to problems of long-term development of regions and systems of regions. "Long-term" basically means that the focus is on structural rather than marginal changes of regional economies. Some of these structural issues of regional policy making are analyzed more efficiently with a continuous two-dimensional spatial representation than with a discrete subdivision into regions, which is more common in regional economics.

To stimulate qualitative, structural analysis of regional development issues, IIASA invited two experts, Martin J. Beckmann and Tönu Puu, to work together at Laxenburg for a short period in September 1979. Time has not yet permitted the authors to complete their monograph, though some of the finished chapters have been circulated as IIASA Collaborative Papers.

One outcome of this work, however, is the present paper by Tönu Puu. It dates from the same period, but was completed when he visited IIASA again in March 1982.

BORIS ISSAEV†
Leader
Regional and Urban Development Group

CONTENTS

STRUCTURALLY STABLE TRANSPORT FLOWS AND PATTERNS OF LOCATION

Tönu Puu
Department of Economics, Umeå University, S-901 87 Umeå, Sweden

SUMMARY

This report describes developments of the continuous model of trade and equilibrium in two-dimensional space, introduced by Martin J. Beckmann in the early 1950s. The model has two distinctive features:

1. *An optimum flow field is found by solving a variational problem for an isotropic metric of transportation cost. The condition is analogous to Huygens' principle in optics.*
2. *There is a connection between the local change in density of commodity flow (its divergence) and the source–sink distribution. This corresponds to the conservation equation in hydrodynamics.*

Except for its capability of representing very general geometries of spatial economies, the Beckmann model is shown to be ideal for the application of structural stability analysis.

The original model is extended by treating several interrelated commodity flows and an explicit production activity, transforming the contents of one flow (labor services) into another (finished goods). A residential–industrial agglomeration pattern arises that corresponds to the two flows.

This general model, which is capable of representing very diverse spatial organizations but at the same time contains very little information, is specified by using the generic theory of differential equations. Therefore, if structural stability of the flows of commodities is assumed, it is possible to obtain a rather precise topological characterization of the stable flow and of the corresponding spatial organization.

Structural stability implies that:

a. the flow is regular (or topologically equivalent to parallel straight lines) except at a finite number of isolated singularities;
b. these singularities are sources, sinks, or saddle points; and
c. no trajectory joins saddle points.

By assuming these conditions, it is possible to construct (up to topological equivalence) the global graph of stable flow. This graph corresponds to a quadratic (not hexagonal) spatial organization.

On the contrary, it is seen that traditional market area theory, as developed by Christaller and Lösch for homogeneous space, becomes structurally extremely unstable if it is transferred to inhomogeneous space.

The main conclusion is that extreme care should be taken when deriving the results of classical market area theory from nonlinear models. The classical theory is linear and, therefore, always structurally stable. Without linearity (i.e. homogeneous space) stability is no longer guaranteed, but must be expressly assumed. The conclusions about basic spatial organization then become very different.

1 REGIONAL MODELING AND STRUCTURAL STABILITY

Many spatial organization patterns in regional modeling may represent equilibria or optima. If a location structure for economic activity is given, the problem of finding an optimum transportation system can be reasonably well defined, and, conversely, location problems can be solved if transportation possibilities are known. What can we say, however, about concentrating economic activity compared with decentralization, or about different agglomeration patterns when we always suppose that an appropriate transportation system is chosen?

One approach to this problem would be as follows. A location pattern and a transportation system would give rise to a system of flows of goods and services between locations in the region being studied. As partial changes in the location of economic activities occur constantly it is reasonable to suppose that the flow patterns are subject to perturbations. Of particular interest would be flows that are stable to perturbation, i.e. that respond with small changes to small perturbations but keep their qualitative structures, and location patterns that are compatible with such flows. We would expect actual location patterns to be compatible with structurally stable flows most of the time, and that there would be sudden changes whenever the structures become unstable.

This philosophy is, of course, very much influenced by catastrophe theory and considerations of transversality that lie behind it. The relevant mathematical concepts date back to work on generic singularities and structurally stable dynamic systems (Morse 1934, Smale 1967, Thom 1969, Peixoto 1973). However, only with the popularity of applications of catastrophe theory (in almost trivial ways) to various discontinuous changes has the substance of the mathematics behind folds, swallowtails, and butterflies been made accessible to nonmathematicians (Poston and Stewart 1978). This could explain why the powerful concepts of structural stability have not been used much by economists until recently.

2 THE CONTINUOUS MODEL OF TRANSPORTATION

The distinctive feature of the present study, beside the use of structural stability, is the continuous transportation model, which was formulated by Beckmann (1952, 1953)

and Kantorovich (1958). The Beckmann formulation is more appropriate for the present purpose.

The model has two characteristics. Firstly, the transportation system is not specified in terms of a network with sources and sinks at a number of nodes, but in terms of a cost of transfer across each point of the region. For simplicity this cost field can be made isotropic by assuming the cost of transfer to be independent of direction. Optimum routes can be found by variational calculus and correspond to Fermat's principle of least time in geometric optics. Later contributions have developed the approach by mapping the derived routes as geodesics on to a flat (Wardrop 1969) or curved (Angel and Hyman 1970, 1972, 1976) manifold (also Puu 1978).

Secondly, a continuous distribution of point sources and sinks over the region is related to excess supplies and demands. The flow of goods becomes analogous to a flow in hydrodynamics and an important equation, corresponding to the continuity equation, relates the divergence of the vector field representing flow to excess supply at each location. This feature of Beckmann's model seems not to have been developed further and hence the continuous space market theory that is occasionally mentioned in textbooks still has the shape in which Beckmann cast it fifteen years ago. In particular, the theory concerns one kind of good or a number of different goods that are *unrelated*, in that they do not constitute inputs and outputs in some production process, for example.

In this study production activity is explicitly introduced by a Cobb–Douglas production function that applies at all locations. Local supplies of raw materials and other variations of productivity are summarized in the variation of a multiplicative factor. One good is produced by means of three classical production factors: land, capital, and labor. By use of this technology, the inputs can be made substitutable in production.

Land is, of course, immobile and so is capital in the sense of buildings and machinery. For labor and goods it is possible that the distribution of residences differs from the employment distribution and that the consumption distribution differs from the production distribution. The result is flows of goods and labor services that are in general oppositely directed. Related to these flows are price and wage gradients.

Hence transportation possibilities decide wage and price structures. In contrast to wages and prices, capital rent is assumed to be constant in equilibrium, as capital services need not be transported. Land rent, finally, is determined as a residual and may vary over the region. The actual short-run possibilities of production will naturally vary because of the different amounts of capital invested, even if we disregard nonsystematic variations.

A consequence of the variation of wages and prices is that the factor productivities and the choice of labor-intensive or capital-intensive techniques will vary over the region.

3 THE PRODUCTIVE SYSTEM

Let us suppose that a Cobb–Douglas production function is applicable at any location of the economy. A homogeneous product Q is generated from land A, capital K, and labor L. Hence

$$Q = bK^{\alpha}L^{\beta}A^{\gamma} \tag{1}$$

where

$$\alpha + \beta + \gamma = 1 \tag{2}$$

The production function was chosen for its simplicity, but it is not impossible to defend the choice on grounds of realism. The substitutability between capital and labor is a standard assumption in economics. In the economics of agriculture, capital increases the productivity of land and more labor does the same. This accounts for the two other substitutions. An analogous argument holds for industrial production, since a factory can be built in several storeys to reduce the need for land and since lack of space certainly reduces the productivity of labor.

Let the local price of the good be p, land rent g, capital rent r, and wage w. Optimum production requires that

$$\alpha Q/K = r/p \tag{3}$$

$$\beta Q/L = w/p \tag{4}$$

These conditions determine how the ratios of rent and wage to price determine total capital and labor productivities. The production function in (2) can be rewritten as $Q/A = b(K/A)^{\alpha}(L/A)^{\beta}$. This together with (3) and (4), where the left-hand sides can be substituted by $\alpha Q/A\,/(K/A)$ and $\beta Q/A\,/(L/A)$, yields three equations in capital, labor, and output per unit area of land, which hence can be solved when the relevant price ratios are known.

The share of land in the total revenue is $pQ - rK - wL$, which, according to (2–4), equals γpQ. Land rent per unit land area is hence determined by

$$\gamma Q/A = g/p \tag{5}$$

This expression is similar to (3) and (4) but it is no optimum condition. Rather, it expresses how land rent is determined from productive activity per unit land area and from product price at the location.

Finally, (2–5) yield

$$gA + rK + wL = pQ \tag{6}$$

or the result that total revenue is divided between incomes of landowners, capitalists, and workers. This will be of use later when general equilibrium is discussed.

The symbols for production and use of factors of production referred to some profit-maximizing unit firm using a finite portion of land as one input. We now wish to consider a continuum of firms at each point; their varying use of land is reflected in varying densities of production and uses of inputs per unit land area. Therefore, $q = Q/A$, $k = K/A$, and $l = L/A$ are introduced. Equations (1–6) can be rewritten in terms of these new symbols, but making q, k, l, and b functions of the space coordinates (x,y) implies that the change is not purely formal but represents a limiting process.

The production function (1) can now be expressed as

$$q = b\,k^{\alpha} l^{\beta} \tag{7}$$

where, by virtue of (2), $\alpha + \beta < 1$. The optimum conditions for capital (3) and labor (4) become

$$\alpha q/k = r/p \tag{8}$$

$$\beta q/l = w/p \tag{9}$$

The equation determining land rent becomes

$$\gamma q = g/p \tag{10}$$

and the profit exhaustion condition (6) is now

$$g + rk + wl = pq \tag{11}$$

We can see that the price ratios determine the quantities of inputs employed and outputs produced per unit land area by substituting from (8) and (9) into (7):

$$q = \alpha^{\alpha/\gamma} \beta^{\beta/\gamma} (p/r)^{\alpha/\gamma} (p/w)^{\beta/\gamma} b$$

Substituting back into (8) and (9), respectively, produces

$$k = \alpha^{1+\alpha/\gamma} \beta^{\beta/\gamma} (p/r)^{1+\alpha/\gamma} (p/w)^{\beta/\gamma} b$$

$$l = \alpha^{\alpha/\gamma} \beta^{1+\beta/\gamma} (p/r)^{\alpha/\gamma} (p/w)^{1+\beta/\gamma} b$$

which establish the assertion. If we now take (10) into consideration, then

$$\alpha^{\alpha} \beta^{\beta} \gamma^{\gamma} p = r^{\alpha} w^{\beta} g^{\gamma}/b$$

which links together the four prices in the model. As we shall see, the costs of transportation and the optimum flow directions determine the spatial variation of product price and wage rate. We shall also assume that capital is optimally allocated so that capital rent is constant. Thus the last relation determines the land rent at every location.

We can consider the implication of this. From the production side, with given transportation possibilities the whole price structure is completely determined. On the other hand, the utility of various locations for a household depends on residential space available and on consumption of goods. The trade-off between housing space and consumption, however, depends on the four prices, as the cost of housing depends on land and capital rents, income depends on local wages, and consumption possibilities are determined by

the local price level. Accordingly, the highest obtainable utility at a given location is determined by the four prices. It would be a very unlikely coincidence, however, if the stipulation of a constant utility for all locations were automatically fulfilled by the price system determined from the production side.

The conclusion is that if we do not admit that capital rent may vary over the region, it is impossible to make all locations equivalent as residences. Therefore, an optimum allocation of capital, having the same yield everywhere, is incompatible with a spatial structure where residential locations are equivalent. Either a variation of the yield on capital or a variation of residential attractiveness is necessary because of a lack of degrees of freedom. Consequently, either migration flow or a relocation of capital in its accumulation process seems to appear whenever this is not precluded by scarcities of space.

We have not considered explicitly the use of land for housing, but doing this would only add a new demand component for land without increasing the number of degrees of freedom.

Our assumption of a linearly homogeneous production function may seem unduly restrictive. It must be admitted that this is crucial to our analysis, as the areal density of output would not be a well defined function of the areal densities of inputs without linear homogeneity. Dividing (1) by A makes land input disappear only when the exponents add up to unity. (The form of a Cobb–Douglas function, on the other hand, is not essential. We could do equally well with other linearly homogeneous functions.)

How restrictive is the assumption of linear homogeneity? Frisch (1965) argues that variable returns to scale are mainly due to incomplete specification of the inputs (and other factors that influence output). Once everything relevant is listed, we could consider any proportionate changes of scale of process operation as possible. We have by no means a very extensive list of inputs. On the other hand, we deal with very broad categories of inputs and output. We have in mind a process by which land is used as space and as a source of an almost freely available raw material (like a mineral or a biological substance). By the application of services of labor and capital (a produced means of production) the materials are turned into finished goods available for general consumption. With such a heavy aggregation the linear homogeneity does not seem too restrictive an assumption.

In regional science increasing returns are often assumed, in fact much more often than in general economics. This indicates that the assumption reflects a wish to establish certain results about agglomeration. It need not rest on a compelling conviction that increasing returns must be assumed as soon as production theory is applied in a spatial context.

The reasoning about increasing returns, or externalities, or both, in spatial economics often indicates that they serve as proxies for accessibility between productive activities that need much interaction. However, once we account separately for the communications, and hence for the accessibilities, in our system, there seems to be no need for an extra assumption about increasing returns or externalities. Only the purely technological reasons remain, but they are no stronger than in general economics so it seems that we should be allowed to disregard variable returns to scale at the present level of abstraction.

4 FLOWS OF GOODS AND LABOR

Whereas $q(x,y)$ and $l(x,y)$ denote the quantities of product supplied and labor demanded at each location, $q'(x,y)$ and $l'(x,y)$ denote the quantities of product demanded

and labor supplied. As the former are determined by the spatial organization of productive activities the latter depend on the residential location structure. Consumption is, of course, not only out of labor incomes but out of land and capital rents as well. If residential areas tend to be differentiated from industrial areas, excess demand and excess supply distributions will arise such that there is excess supply of goods and excess demand for labor in industrial areas and the opposite situation in residential areas. The differentiation is not such that production exclusively takes place in industrial areas; it is just more concentrated there. As a consequence, flows of goods and labor in opposite directions arise.

We denote the flows of goods and labor, respectively, by

$$\boldsymbol{q} = (q_1(x,y), q_2(x,y)) \tag{12}$$

and

$$\boldsymbol{l} = (l_1(x,y), l_2(x,y)) \tag{13}$$

They are vector fields, i.e. they have direction as well as magnitude, and vary in direction and/or magnitude from one point to another. Conceptual flows of capital and land might be introduced as well by assigning residential location to capitalists and landlords, or by assuming an ownership structure for capital and land among workers. However, it only complicates analysis to introduce these zero-cost flows, which can take any paths in space. The cost is zero because capital and land are already invested at the points of employment.

The unit vectors

$$\boldsymbol{q}/|\boldsymbol{q}| = (\cos\theta, \sin\theta) \tag{14}$$

$$\boldsymbol{l}/|\boldsymbol{l}| = -(\cos\theta, \sin\theta) \tag{15}$$

define the flow directions, which are assumed to be opposite because we assume that the same transportation system is used for goods and labor and because we assume goods to flow from industry to residences and labor services from residences to industry.

The Euclidean norms

$$|\boldsymbol{q}| = (q_1^2 + q_2^2)^{1/2} \tag{16}$$

$$|\boldsymbol{l}| = (l_1^2 + l_2^2)^{1/2} \tag{17}$$

represent quantities of goods and labor shipped across a given location.

The divergences of the flows, $\nabla . \boldsymbol{q} = \partial q_1/\partial x + \partial q_2/\partial y$ and $\nabla . \boldsymbol{l} = \partial l_1/\partial x + \partial l_2/\partial y$, represent the quantities of elements added to the flows at each point if positive or the quantities withdrawn if negative. According to the related vector theorems of Gauss, Green, and Stokes the divergence of a vector field that represents a flow equals a source or negative sink density (Marsden and Tromba 1976). Hence we can equate divergence with excess supply:

$$\nabla . \boldsymbol{q} = q - q' \tag{18}$$

$$\nabla_{\bullet} \boldsymbol{l} = l' - l \tag{19}$$

Even though everything could be expressed in terms of partial derivatives, the use of the nabla operator $\nabla_{\bullet}$ saves a lot of writing. Equations (18) and (19) are the continuity equations referred to in Section 2.

5 OPTIMUM FLOW LINES

The next task is to determine the optimum flow lines, given a system of transportation possibilities. As said in the introduction, we are not going to specify any network. Instead we assume a function

$$f(x,y) \tag{20}$$

that determines the cost of transfer or displacement of some quantum of goods or labor across the point (x,y). Since the cost field does not depend on the direction θ of passage we deal with an isotropic problem. Without loss of generality we can fix the units so that f is the transfer cost both for one unit of goods and for one unit of labor.

If we have any parametrized curve $x(\sigma)$, $y(\sigma)$, where σ is the "natural" arc length parameter, we can define the cost of transportation of one quantum of goods or labor over a distance s by the path integral

$$c(s) = \int_0^s f \, \mathrm{d}\sigma \tag{21}$$

If we fix two endpoints by the boundary conditions $x(0) = x_0$, $y(0) = y_0$ and $x(s) = x_1$, $y(s) = y_1$, the value of c depends on the choice of the arc connecting the endpoints. This yields a well defined variational problem as the transporters would seek to minimize (21) for each pair of endpoints. The variational problem is solved by the appropriate Euler equation (Fox 1954). The solution to the relevant differential equation is really an extremum (Puu 1978) because the Jacobi and Legendre conditions are fulfilled.

A more elegant way of formulating the problem is to find vector fields $\boldsymbol{q}$ and $\boldsymbol{l}$ that, subject to the constraints (18) and (19), minimize total transportation costs for goods and labor over the region $\mathcal{R}$:

$$T_q = \iint_{\mathcal{R}} f \, |\boldsymbol{q}| \, \mathrm{d}x \, \mathrm{d}y \tag{22}$$

$$T_l = \iint_{\mathcal{R}} f \, |\boldsymbol{l}| \, \mathrm{d}x \, \mathrm{d}y \tag{23}$$

This formulation was used by Beckmann (1952, 1953) and is intuitively plausible. However, these cost expressions can be transformed in a reasonable way from fourfold integrals of the product of c according to (21) and the quantity shipped for each pair of endpoints (Puu 1977).

The constrained optimization problem for vector fields can be transformed into an unconstrained problem of minimizing double integrals, over the region $\mathcal{R}$, of the integrands in (22) and (23), to which the constraints (18) and (19) multiplied by a Lagrangian function $\lambda(x,y)$ are added. We can use the same multiplier function because we have assumed the flows to have opposite directions. The signs are, of course, reversed. This yields the expressions to be minimized:

$$\iint_{\mathcal{R}} [f|\boldsymbol{q}| - \lambda(q - q' - \nabla \cdot \boldsymbol{q})]\, dx\, dy \tag{24}$$

$$\iint_{\mathcal{R}} [f|\boldsymbol{l}| - \lambda(l - l' + \nabla \cdot \boldsymbol{l})]\, dx\, dy \tag{25}$$

For minima the Euler equations

$$f\boldsymbol{q}/|\boldsymbol{q}| = \nabla\lambda \tag{26}$$

$$f\boldsymbol{l}/|\boldsymbol{l}| = -\nabla\lambda \tag{27}$$

must be satisfied. As usual $\nabla\lambda$ denotes the gradient vector (λ_x, λ_y). The left-hand sides of (26) and (27) are hence oppositely directed potential flows. The flows $\boldsymbol{q}$ and $\boldsymbol{l}$ are co-directional, but they need not be potential flows themselves, owing to the multiplicative factors. Hence the actual flows of goods and labor may be rotational, but the fact that their flow lines are obtainable from a potential is an important conclusion because it rules out certain types of critical points (spirals and centers). We shall return to this later.

Let us now multiply both sides of (26) and (27) by the unit vectors $\boldsymbol{q}/|\boldsymbol{q}|$ and $\boldsymbol{l}/|\boldsymbol{l}|$, respectively. The product of two identical unit vectors being a scalar unitary number, the left-hand sides equal f. On the other hand, these unit vectors are in the directions of the path and thus their products with the gradient of λ equal $\pm\, d\lambda/ds$. However, we see from (21) that f is the increment of transportation cost along the path. If the price and the wage increase with transportation cost along the optimum paths, which is an obvious assumption, then

$$dp/ds = d\lambda/ds \tag{28}$$

$$dw/ds = -d\lambda/ds \tag{29}$$

Integrating, we obtain

$$p = \bar{p} + \lambda \tag{30}$$

$$w = \bar{w} - \lambda \tag{31}$$

Hence $\lambda(x,y)$ = constant represents a family of coincident curves in space for constant price and constant wage. Since, according to (26) and (27), the flow lines for goods and labor are coincident with the gradient direction (or its opposite) we can conclude that the lines of constant price and wage are cut orthogonally by the (oppositely directed) flow lines for goods and labor.

The potential function $\lambda(x,y)$ hence contains much information and we can expect that it will play an important role later.

6 EXAMPLES OF OPTIMUM PATHS

To give more substance to the discussion we can examine the system of optimum routes for some specific class of transfer cost functions. We assume a function with circular symmetry with respect to the origin. This is a natural choice, having connections back to von Thünen and with the literature of the "New Urban Economics" in which circular region shapes and other kinds of circular symmetry are abundant.

We hence write $f(\rho)$ where $\rho = (x^2 + y^2)^{1/2}$. The simplest way of obtaining the relevant Euler equation is to start with (21), substituting the arc length element $d\sigma$ with $[\rho^2 + (d\rho/d\omega)^2]^{1/2}\, d\omega$. It is natural to use polar coordinates, where $\rho = (x^2 + y^2)^{1/2}$ and $\omega = \tan^{-1}(y/x)$, the latter being treated as an independent variable. The Euler equation is then

$$f\rho^2/[\rho^2 + (d\rho/d\omega)^2]^{1/2} = \text{constant} \tag{32}$$

If, for illustrative simplicity, we now specify the transfer cost function as

$$f = \rho^{n-1} \tag{33}$$

then the solution is

$$\rho^n = a \sec(n\omega + b) \tag{34}$$

unless $n = 0$.

We shall study a few cases of special interest, transforming the solution formulae back from polar to Cartesian coordinates. For $n = 2$ and $n = 3$,

$$(x^2 - y^2)\cos b - 2xy \sin b = a \tag{35}$$

$$(x^3 - 3xy^2)\cos b - (3x^2y - y^3)\sin b = a \tag{36}$$

which represent families of level curves for an ordinary saddle and a so-called monkey saddle, respectively. Different b for a given a only represent various rotations of some basic curve of given shape around the origin, whereas different a for a given b yield a whole family of nonintersecting curves that covers the whole plane. Thus nothing substantial is lost by assuming $b = 0$, so that (35) and (36) become $x^2 - y^2 = a$ and $x^3 - 3xy^2 = a$. The solutions are illustrated in these forms in Figures 1 and 2.

For $n = -2$,

$$(x^2 - y^2)\cos b - 2xy \sin b = a(x^2 + y^2)^2 \tag{37}$$

which yields a family of lemniscates. Again, we lose nothing by assuming a certain rotation of the curve system by putting $b = 0$, so that $x^2 - y^2 = a(x^2 + y^2)^2$ gives a family for

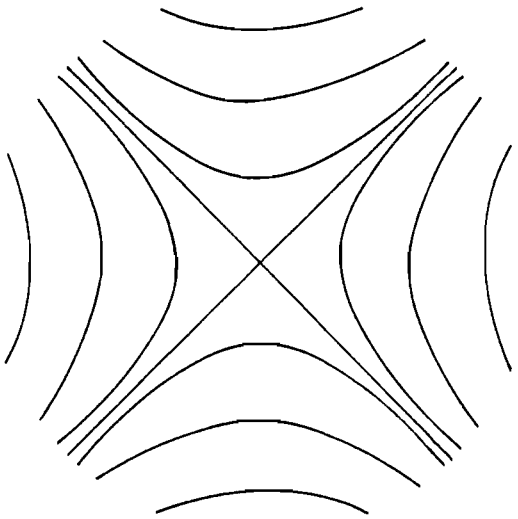


FIGURE 1 Saddle flow.

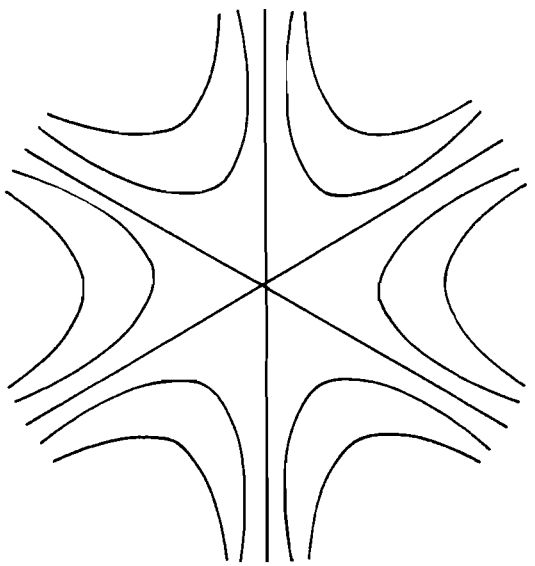


FIGURE 2 Monkey saddle flow.

various a that covers the whole plane. The curves intersect only at the origin. The solution is shown in Figure 3.

Finally, the simplest cases are $n = 1$ or -1:

$$x \cos b - y \sin b = a \tag{38}$$

$$x \cos b - y \sin b = a(x^2 + y^2) \tag{39}$$

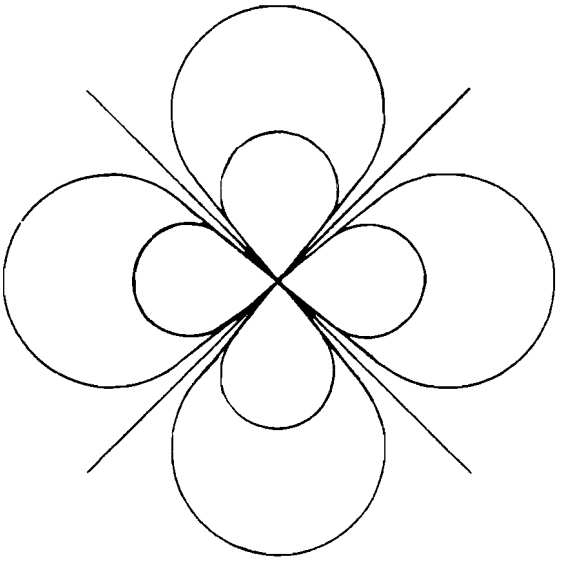


FIGURE 3 Lemniscate flow.

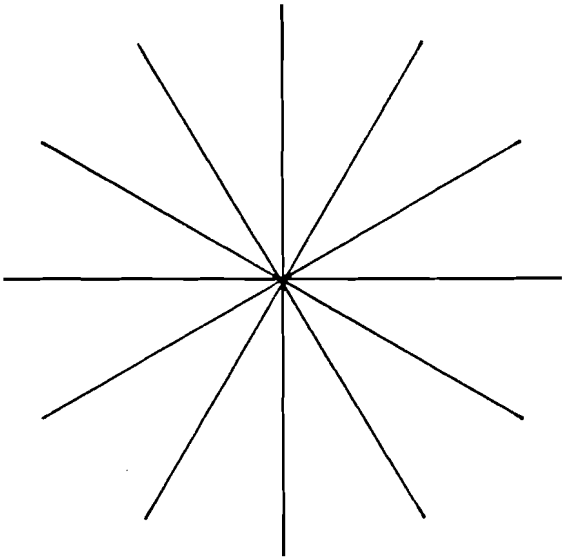


FIGURE 4 Radial flow.

are the solutions. They represent families of straight lines and circular segments, respectively. As before, we are interested in patterns with some rotational group symmetry and such that the families cover the whole plane with curves that do not intersect (with the origin as a possible exception). The only way to obtain such patterns from (38) and (39) is to put $a = 0$, while b varies to generate the families. Accordingly, both (38) and (39) become $y/x = \tan b$, which represents a family of radials. Although it may seem surprising

that the circles turn into straight lines, they are just circles of infinite radius. The case is illustrated in Figure 4.

The orthogonal price–wage curves corresponding to the illustrated flows are not difficult to obtain. In Figure 4 they would be a family of concentric circles, whereas in Figures 1, 2, and 3 they look like the trajectory families themselves but rotated by an angle of 30–45°. The four types of flow illustrated will be of interest later in this article. Running ahead of argument, we shall see that Figure 3 and the case of Figure 4 where the radii are "circles" are excluded because the flows actually intersect at the origin. There may be a confluence of flows that stagnate at the origin, as the other case illustrated in Figure 4 demonstrates, but no crossing. We shall see also that cases like that in Figure 2 are excluded because the pattern is structurally unstable and the singular point at the origin may be suddenly split by the smallest perturbation.

7 GENERAL EQUILIBRIUM PROPERTIES OF THE MODEL

We shall next study the general equilibrium properties of the model to check its internal consistency. Therefore, we shall study the *value* flows pq for goods and wl for labor and their divergences in particular.

One of the fundamental tools in vector analysis is a theorem that makes the surface integral of the divergence of any regular vector field on some bounded region equal to the line integral of the outward component of the field along the boundary. The theorem is usually called Gauss's theorem or the divergence theorem. Its general form is relevant for a surface in three-dimensional space, but our interest is in a much simpler version for a two-dimensional plane. We have already touched on the subject in the comments on the interpretation of a divergence. Obviously, if we shrink the bounded region to a single point then the line integral gives net outflow from the single point. To unveil some of the mystery about the theorem we could also stress that it has a parallel with the fundamental theorem of calculus that relates the value of a definite integral to the values of the primitive function on the boundary. Thus

$$\iint_{\mathcal{R}} \nabla . (\boldsymbol{pq})\, \mathrm{d}x\, \mathrm{d}y = \oint_{\partial \mathcal{R}} \boldsymbol{pq} . \boldsymbol{n}\, \mathrm{d}s \tag{40}$$

$$\iint_{\mathcal{R}} \nabla . (\boldsymbol{wl})\, \mathrm{d}x\, \mathrm{d}y = \oint_{\partial \mathcal{R}} \boldsymbol{wl} . \boldsymbol{n}\, \mathrm{d}s \tag{41}$$

are the formulae we need. Here $\boldsymbol{n}$ is the usual symbol for a unit outward vector normal to the boundary $\partial \mathcal{R}$ and the line integrals on the right-hand sides are taken in a positive direction (i.e. the one that leaves the interior of $\mathcal{R}$ to the left).

Since these line integrals denote net value outflow from the region they must equal the differences between the values of export and import. By introducing the symbols X_q, X_l for the value exports of goods and labor and M_q, M_l for the value imports of goods and labor, we can write

$$\oint_{\partial \mathcal{R}} \boldsymbol{pq} . \boldsymbol{n}\, \mathrm{d}s = X_q - M_q \tag{42}$$

$$\oint_{\partial R} w\boldsymbol{l}.\boldsymbol{n}\, ds = X_l - M_l \tag{43}$$

Next we note a theorem from vector analysis by which the divergence of the product of a scalar and a vector equals the dot product of the gradient of the scalar and the vector plus the product of the scalar and the divergence of the vector. Marsden and Tromba (1976) may be consulted about this, as well as about Gauss's theorem. Hence we can expand the left-hand sides of (40) and (41) as

$$\iint_R \nabla_{\circ}(p\boldsymbol{q})\, dx\, dy = \iint_R \nabla p.\boldsymbol{q}\, dx\, dy + \iint_R p\nabla.\boldsymbol{q}\, dx\, dy \tag{44}$$

$$\iint_R \nabla_{\circ}(w\boldsymbol{l})\, dx\, dy = \iint_R \nabla w.\boldsymbol{l}\, dx\, dy + \iint_R w\nabla.\boldsymbol{l}\, dx\, dy \tag{45}$$

How can the right-hand sides of (44) and (45) be interpreted? First, we note from (30) and (31) that $\nabla p = \nabla\lambda$ and $\nabla w = -\nabla\lambda$. Thereafter we see from (26) and (27) that $\nabla\lambda.\boldsymbol{q} = f|\boldsymbol{q}|$ and $-\nabla\lambda.\boldsymbol{l} = f|\boldsymbol{l}|$. This is so because $\boldsymbol{q}.\boldsymbol{q} = |\boldsymbol{q}|^2$ and $\boldsymbol{l}.\boldsymbol{l} = |\boldsymbol{l}|^2$. Collecting the results, we can write

$$\iint_R \nabla p.\boldsymbol{q}\, dx\, dy = \iint_R f|\boldsymbol{q}|\, dx\, dy \tag{46}$$

$$\iint_R \nabla w.\boldsymbol{l}\, dx\, dy = \iint_R f|\boldsymbol{l}|\, dx\, dy \tag{47}$$

However, according to (22) and (23) these right-hand sides equal the transportation costs for goods and labor respectively, T_q and T_l.

As to the interpretation of the second terms on the right-hand sides of (44) and (45), we see from (18) and (19) that

$$\iint_R p\nabla.\boldsymbol{q}\, dx\, dy = \iint_R p(q - q')\, dx\, dy \tag{48}$$

$$\iint_R w\nabla_{\circ}\boldsymbol{l}\, dx\, dy = -\iint_R w(l - l')\, dx\, dy \tag{49}$$

We are now able to collect the results. Substitution from (42), (44), (46), (22), and (48) into (40) and from (43), (45), (47), (23), and (49) into (41) yields

$$\iint_R p(q - q')\, dx\, dy = X_q - M_q - T_q \tag{50}$$

$$\iint_R w(l - l')\, dx\, dy = -(X_l - M_l - T_l) \tag{51}$$

Equation (50) shows that the value of all excess supplies of goods evaluated at the local price levels equals the value of exports of goods minus the value of imports of goods minus

the cost of transportation for goods. Likewise, according to eqn. (51) the value of all excess demands of labor evaluated at the local wage rates equals the negative of the value of exports of labor services minus the value of imports of labor services minus the cost of transportation for labor. That transportation cost is equivalent to imports is not at all surprising as transportation services are in a certain sense imported because we have not accounted for any use of inputs for transportation. It would be easy to change this fact by introducing some transportation technology in terms of some simple production function for transportation services, but this only complicates the model without yielding more than obvious conclusions.

As already noted, land and capital are treated differently from goods and labor. This is so because land and capital are already at the place of employment, whereas labor services and goods produced need to be transported at nonzero freight rates. By this procedure we miss the possibility of specifying a spatial ownership structure for land and capital. However, differentiating between demand and supply of land and capital, where the former depends on the production structure and the latter on the ownership structure, would double the number of derivations, (40–51), again without yielding more than obvious conclusions. Therefore, incomes from land rent and capital rent need not be referred to any spatial structure because purely monetary transfers in space are assumed to be costless.

We define total income from land rent as

$$G = \int\int_{\mathcal{R}} g \, dx \, dy \tag{52}$$

and total income from capital rent as

$$R = \int\int_{\mathcal{R}} r k \, dx \, dy \tag{53}$$

Further we define labor incomes as

$$W = \int\int_{\mathcal{R}} w l' \, dx \, dy \tag{54}$$

and total consumption as

$$C = \int\int_{\mathcal{R}} p q' \, dx \, dy \tag{55}$$

If we now integrate both sides of eqn. (11) over the region and substitute from eqns. (50–55), we obtain

$$G + R + W - C = X - M - T \tag{56}$$

We have lumped exports, imports, and transportation costs together so that $X = X_q + X_l$, $M = M_q + M_l$, and $T = T_q + T_l$. The equation states that the sum of all incomes from land rent, capital rent, and wages minus regional consumption equals exports minus imports minus the costs for transportation that are assumed as imported. This establishes

the consistency of the system conceived as a general equilibrium. For simplicity we do not consider flows of incomes from land and capital ownership across the boundary. This is another generalization that would be easy to put into effect but it would make the formalism grow disproportionately to the extra conclusions that could be drawn.

8 OPTIMALITY

We have treated the problem as a competitive equilibrium with individually optimizing agents. The natural question arises whether this equilibrium represents a social optimum or not. To answer this we take a planning model as illustration, where the welfare of the inhabitants of an isolated area is maximized within the limits imposed by the available resources. The optimality conditions include those for individual producers and transporters as stated above. We shall also use this discussion to show how transportation services can be made endogenous.

Let us suppose that we wish to maximize

$$\iint_{\mathcal{R}} l' U(q'/l') \, dx \, dy \tag{57}$$

where $U(q'/l')$ is the individual utility of per capita consumption. At each location we multiply individual utility by the population and assume that utilities are additive over locations. As before, local production depends on inputs according to

$$q = b k^{\alpha} l^{\beta} \tag{58}$$

and the excess of local production above local consumption enters the flow as before, i.e.

$$q - q' = \nabla . \boldsymbol{q} \tag{59}$$

We now suppose that transportation uses up κ units of capital and λ units of labor per unit flow density. These fixed coefficients are assumed to be functions of the space coordinates. As capital and labor are now needed for transportation the constraints can be expressed as

$$\iint_{\mathcal{R}} [k + \kappa(|\boldsymbol{q}| + |\boldsymbol{l}|)] \, dx \, dy = K \tag{60}$$

and

$$l' - l - \lambda(|\boldsymbol{q}| + |\boldsymbol{l}|) = \nabla . \boldsymbol{l} \tag{61}$$

The constraints look different in two respects. The constraint for capital is in integral form as we are free to choose the spatial distribution of capital, whereas the constraint for labor is in local form as the supply of labor is given at each location. Moreover, there

may be local discrepancies between labor supply and demand that are equal to the divergence of the flow. This is not so with capital.

We now associate Lagrangian multipliers p, r, and w with the constraints (59), (60), and (61) and are ready to formulate an optimization problem. We maximize (57) with respect to consumption, capital, and labor used in production at every location and with respect to the flows of goods and labor subject to the constraints, including (58), which is substituted into (59). The optimum conditions are

$$\mathrm{d}U/\mathrm{d}(q'/l') = p \tag{62}$$

$$\alpha q/k = r/p \tag{63}$$

$$\beta q/l = w/p \tag{64}$$

$$(r\kappa + w\lambda)\boldsymbol{q}/|\boldsymbol{q}| = \nabla p \tag{65}$$

$$(r\kappa + w\lambda)\boldsymbol{l}/|\boldsymbol{l}| = \nabla w \tag{66}$$

Among these equations (63) and (64) are familiar as conditions for maximizing the profits of individual producers. From the integral form of (60) we conclude that the corresponding Lagrangian multiplier r is a constant over locations. The Lagrangian multipliers also receive the interpretations of production and factor prices. If we define profits as being $g = pq - rk - wl = (1 - \alpha - \beta)pq = \gamma pq$ we again have

$$\gamma q = g/p \tag{67}$$

where we can interpret the profits as land rent.

The expression $r\kappa + w\lambda$ obviously denotes the local displacement cost, and we can again denote it by f if we wish. The cost of displacement is the same in (65) and (66), so the two gradients have the same modulus and we can define $\nabla p = \nabla\lambda$ and $\nabla w = -\nabla\lambda$. Again, goods and labor flow in the directions of steepest increase of prices and wages, which increase with transportation cost in those directions. The only new condition is (62), stating that marginal utility must equal product price everywhere.

By specifying the need for transportation of inputs we have endogenized transportation. We can show this by multiplying (65) and (66) by $\boldsymbol{q}$ and $\boldsymbol{l}$, respectively:

$$(r\kappa + w\lambda)|\boldsymbol{q}| = \nabla p \,.\, \boldsymbol{q} \tag{68}$$

$$(r\kappa + w\lambda)|\boldsymbol{l}| = \nabla w \,.\, \boldsymbol{l} \tag{69}$$

As there is no trade at all in goods or labor between the isolated area and the outside we conclude from Gauss's theorem that the integrals of the divergences of both value flows, pq and wl, on $\mathfrak{R}$ must equal zero. Hence

$$\iint_{\mathfrak{R}} \nabla p \,.\, \boldsymbol{q}\, \mathrm{d}x\, \mathrm{d}y + \iint_{\mathfrak{R}} p\, \nabla \,.\, \boldsymbol{q}\, \mathrm{d}x\, \mathrm{d}y = 0 \tag{70}$$

$$\iint_{\mathcal{R}} \nabla w \cdot \boldsymbol{l}\, dx\, dy + \iint_{\mathcal{R}} w \nabla \cdot \boldsymbol{l}\, dx\, dy = 0 \tag{71}$$

Using (68) and (69) along with (59), (60), (61), and the identity $pq = rk + wl + g$ in (70) and (71), we arrive at

$$\iint_{\mathcal{R}} pq'\, dx\, dy = rK + \iint_{\mathcal{R}} wl\, dx\, dy + \iint_{\mathcal{R}} g\, dx\, dy \tag{72}$$

where in particular all the transportation cost expressions have disappeared. The left-hand side equals consumption and the right-hand side the sum of capital, labor, and land incomes. In the symbols used before,

$$C = R + W + G \tag{73}$$

We have thus seen that the competitive equilibrium conditions for autonomous firms and transporters are the same as the social optimum conditions when the welfare of the inhabitants of the region is maximized. In particular, the optimum conditions are independent of the form of the utility function. The only equation in which the latter appears at all is (62). This condition, of course, puts a further restriction on the type of spatial organization, but it in no way conflicts with the other optimality conditions for transportation and production. The same is true about the aggregation conditions, (72) and (73). This is of particular interest as they state that for the whole region an aggregate budget constraint is fulfilled with all quantities evaluated at local prices. This implies that, even if a spontaneous equilibrium solution to the spatial equilibrium problem were not in accordance with (62), it would be possible to design an internal income transfer policy within the region, because the fulfillment of the aggregate budget constraint admits local fulfillment. The conclusion is that the social optimum is also compatible with consumer autonomy, provided that a proper transfer policy is designed.

9 UNIQUENESS OF THE FLOW PATTERNS

In the discussion of Figures 1–4 we mentioned that flow trajectories should not cross. The reason for this is obvious. In Section 5 we showed that the flow lines coincide with the gradient directions of a potential function, so that the level curves of this potential, which represent the loci of constant wage and constant price, are cut orthogonally by the flow lines for goods and labor.

If two sets of flow lines did cut each other then the orthogonal price and wage lines would intersect as well, but what would be the consequence? Since, then, different price–wage curves in one family of level curves would be intersected by any one curve of the intersecting family, prices and wages along the intersecting curve would be at once equal and different. As we suppose a system of competitive pricing only one product price and one wage rate are associated with each location and thus the case of intersecting price–wage curves or, which is the same, of intersecting trajectories leads to a contradiction.

The potential function conceived as a surface would have a very curious look because there would be self-intersections of the surface whenever the trajectories intersected.

What has been said does not rule out the possibility that there may be points of confluence for the trajectories, provided that the flows *stagnate* at these critical points. The following study will examine which types of such critical points are likely to occur and how these critical points are typically related to each other. As we shall see, these considerations result in one basic type of flow pattern. This, however, does not mean that the flows cross, as they stagnate when the flow lines seem to intersect.

10 ANALYTICITY OF THE POTENTIAL FUNCTION

As has already been stressed, the potential function $\lambda(x,y)$ plays a fundamental role in the discussion. Section 9 led to the conclusion that the potential function is single-valued, because prices and wages are unique functions of the space coordinates. We also discussed critical points where the flows stagnate. As the directions of the flows are determined by the gradient directions to the potential the critical points obviously correspond to stationary points of the potential function. Later we shall show that only elliptic and hyperbolic stationary points are likely, i.e. where the surface has an isolated maximum or minimum or an ordinary saddle, and that it is unlikely that characteristic lines join two different stationary points. This yields the main conclusions and results in a definite basic pattern of the flows.

Before entering these matters we make the simple and very weak assumption that the potential function is analytic, i.e. it can be expanded in a Taylor series in some neighborhood of any point. This assumption makes it easier to work with a complex analytic function. This may seem restrictive as analyticity for a complex function requires fulfillment of the Cauchy–Riemann differential equations, but these are in fact equivalent to the condition that the function may be expanded in a power series that converges. The particular case where the function is real hence only means expandability in a convergent Taylor series. The advantage is that, by working with complex functions, we can make use of the very powerful tools that Cauchy's integral formulae represent. The reader is referred to Marsden (1973) for complex analysis in general, and to Cartan (1963) for the case of an analytic function of two complex variables. The symbols in this section are used in a completely different sense from possible uses in the rest of the study.

A general analytic complex function of two complex arguments z_1, z_2 can be written as a power series:

$$f = \sum_{k=0}^{\infty} \sum_{i+j=k} a_{ij} z_1^i z_2^j \tag{74}$$

The series has been written as a double sum because there are certain advantages in assembling all terms of the same degree in one sum and in making a second summation over all degrees. By using Cauchy's integral formulae, which determine the values of any complex analytic function and its derivatives for any point from its values on a closed curve surrounding this point, we can easily evaluate the coefficients a_{ij} as these depend on the various

partial derivatives. Hence, taking a so-called polydisk where both complex variables vary along circles so chosen that the function converges, we obtain

$$a_{ij} = -\frac{1}{4\pi^2} \iint_{\substack{|z_1| = R_1 \\ |z_2| = R_2}} \frac{f(z_1, z_2)\, dz_1\, dz_2}{z_1^{i+1} z_2^{j+1}} \tag{75}$$

where R_1, R_2 are the radii of the circles along which the two integrals are taken. For simplicity we have assumed that the expansion is made at the origin.

Next we define

$$M = \max_{\substack{|z_1| = R_1 \\ |z_2| = R_2}} |f(z_1, z_2)| \tag{76}$$

Then from (75) and (76),

$$|a_{ij}| \leqslant M/(R_1^i R_2^j) \tag{77}$$

and if $R = \min(R_1, R_2)$,

$$\left| \sum_{i+j=k} a_{ij} z_1^i z_2^j \right| \leqslant M(k+1)(r/R)^k \tag{78}$$

holds true whenever $r = \max(|z_1|, |z_2|)$. Since we suppose that the arguments only take values within the radius of convergence this last condition holds true and, moreover, we have $r/R < 1$ for the same reason. The important relations (77) are called Cauchy's inequalities. The factor $k + 1$ is simply the number of different monomials of degree k.

The relation (78) could be applied equally well to real variables, in which case we put $z_1 = x$, $z_2 = y$. Then r is the absolute value of the two variables x, y that is greater and it is still less than R, for which the series converges.

We now express the series (74) in real arguments and with real coefficients. Let us write the terms of the three lowest orders explicitly. What then can we say about the remainder of the series? The answer is easy to obtain with the help of (78). The remainder of the terms, denoted by Tayl, is, because of (78), subject to the following inequality:

$$\left| \sum_{k=3}^{\infty} \sum_{i+j=k} a_{ij}\, x^i y^j \right| \leqslant M \sum_{k=3}^{\infty} (k+1)\rho^k \tag{79}$$

where we define $\rho = r/R < 1$. The sum on the right-hand side of (79) can be evaluated to $\rho^3 [1/(1-\rho)^2 + 3/(1-\rho)]$. Hence, remembering that $\rho^3 = r^3/R^3$, we have

$$|\text{Tayl}/r^3| \leqslant M [1/(1-\rho)^2 + 3/(1-\rho)]/R^3 \tag{80}$$

which obviously is finite since ρ is less than unity; M is the finite maximum of the function on one of the two circles and R is the positive radius of the smaller of the circles.

If we put $r = (x^2 + y^2)^{1/2}$ instead of equating it to the larger of x and y then the left-hand side of (80) would be diminished even further and the inequality would hold with this new interpretation of r. As r goes to zero so does ρ and hence

$$\lim_{r \to 0} |\,\mathrm{Tayl}/r^3\,| \leqslant 4M/R^3 \tag{81}$$

The right-hand side is finite and hence the ratio of Tayl to r^3 stays finite as r goes to zero. The standard expression for this property is

$$\mathrm{Tayl} = \mathrm{O}(r^3) \tag{82}$$

11 THE CRITICAL POINTS OF THE FLOWS ARE SIMPLE

The result of the preceding section is that if we assume the potential function to be analytic then we can write its Taylor series as a sum of the terms of the three lowest orders with a remainder whose ratio to ρ^3 stays finite as $\rho = (x^2 + y^2)^{1/2}$ approaches zero. We shall now continue with the original notation of this study. Implicit in the discussion of analytic functions was that they were expanded at the origin, $x = y = 0$. We are particularly interested in critical points where the potential function is stationary. Hence it is natural to put the critical point studied at the origin.

At a critical point the conditions for stationarity of the potential function $\lambda(x,y)$ are that the first partial derivatives are equal to zero: $\lambda_x = \lambda_y = 0$ at $x = y = 0$. Moreover, we lose nothing in generality by assuming that $\lambda = 0$ as well at $x = y = 0$. All this only amounts to a translation of the coordinate system to put the critical point at the origin of (x,y,λ) space. This means that at the origin the constant and the linear term vanish. However, since we have decided to write down the terms of the three lowest orders (from zeroth to second) the potential may be expressed as

$$\lambda(x,y) = (\lambda_{xx}x^2 + 2\lambda_{xy}xy + \lambda_{yy}y^2)/2! + \mathrm{O}(r^3) \tag{83}$$

Only quadratic terms are explicit in this expansion. The partial derivatives are constants evaluated at the origin. To simplify notation and avoid misunderstanding, we make the convention that when the potential function or its derivatives are written with explicit arguments they are to be interpreted as functions, whereas λ or its derivatives written without arguments denote the constant values at the origin. This practice is used in this section only and saves some notation.

As we have assumed the potential function to be analytic we can calculate its derivatives by differentiating term by term in the series. This is an elementary result for all analytic functions. We hence obtain

$$\lambda_x(x,y) = \lambda_{xx}x + \lambda_{xy}y + \mathrm{O}(r^2) \tag{84}$$

$$\lambda_y(x,y) = \lambda_{yx}x + \lambda_{yy}y + \mathrm{O}(r^2) \tag{85}$$

near the origin. This illustrates the convention made. The derivatives on the left-hand sides are conceived as functions and those on the right-hand sides as constants. The degree of the remainder terms is obviously lowered by one through differentiation. The fact that the remainder terms are $O(r^2)$ implies that they are $o(r)$, which means that the ratio of the remainder to r goes to zero as r does. This must be true if the ratio of the remainder to r^2 stays finite in the same limiting process.

As the flow lines for goods and labor were seen to coincide in direction with the gradient of $\lambda(x,y)$, i.e. with $\nabla\lambda = (\lambda_x(x,y), \lambda_y(x,y))$ – even though the flows were not necessarily gradient flows – then there exists a parametrization of the flow lines such that $\dot{x} = \lambda_x(x,y)$ and $\dot{y} = \lambda_y(x,y)$. The derivative of a parametrized curve with respect to the parameter will be denoted by a dot. Thus

$$\dot{x} = \lambda_{xx}x + \lambda_{xy}y + o(r) \tag{86}$$

$$\dot{y} = \lambda_{yx}x + \lambda_{yy}y + o(r) \tag{87}$$

are the relevant differential equations in the vicinity of a critical point. These equations differ from linear equations of the simplest form by the $o(r)$ term only. Hence the critical points are what in the theory of ordinary differential equations are called *simple critical points*. A standard result is that simple, isolated critical points look precisely like those of the corresponding purely linear systems (without the $o(r)$ terms). As the latter are well known and classified in a few simple categories the situation is fortunate. The requirement for this is that the nonlinear systems differ from the linear ones only by terms that go to zero faster than the linear terms, and this holds for an analytic potential. Simmons (1972) can be consulted about critical points for differential equations.

It still remains to be shown that the critical points are isolated as well as simple, otherwise the conclusion cannot be drawn. This, however, will be accomplished by using transversality considerations. The result will be that the critical points are nodes, saddles, or spirals. It will, moreover, be demonstrated that spirals (including centers) are ruled out because the flow lines coincide with those of a potential flow, which thus leaves two categories of critical points only.

Before continuing we should consider how restrictive the assumption of analyticity is. All the well known elementary functions are analytic and it is difficult to construct an example of an explicit function as a compound of these elementary functions that is not analytic and renders a system of differential equations that differs from a linear system by more than $o(r)$. The assumption of analyticity is thus not restrictive.

12 TRANSVERSALITY AND MORSE FUNCTIONS

We must now make it probable that critical points are isolated. This will be done by showing that the potential function is a Morse function at a critical point, i.e. it can be transformed to a Morse saddle by some smooth change of coordinates, or, which is the same, that the critical points are either elliptic or hyperbolic in the terminology of differential geometry. This will be accomplished by using transversality considerations that make it unlikely that the Hessian $\lambda_{xx}\lambda_{yy} - \lambda_{xy}^2$ is zero at a critical point where λ_x and

λ_y are zero. At the same time as we rule out whole curves of critical points or areas of them, we also exclude some complicated shapes of isolated critical points like monkey saddles. These points are structurally unstable and split by the slightest deformation of the potential surface.

Transversality is concerned with typical crossings of manifolds in some surrounding space in which they are embedded. It introduces probability considerations in pure mathematics that usually deals only with what is possible and what is impossible. In the Euclidean plane two straight lines through the origin would be very unlikely to coincide if all inclinations are *a priori* equally probable for the lines. The probability measure for coincidence would be zero, even if it were possible. In the natural Euclidean space of three dimensions two planes through the origin would for the same reason typically intersect in a line. The probability of coincidence, so that the intersection is a plane, would again have zero probability measure. If a plane and a line pass through the origin in the same space the typical intersection would be a single point. That the line lies in the plane is unlikely.

That a crossing is likely, in that its probability measure is not zero, is exactly what is meant by a *transverse crossing*. The preceding examples make transversality depend on the dimensions of the intersecting subspaces, of the intersection space, and of the surrounding space. The reasoning has been in terms of linear subspaces, but there is no difficulty in changing the picture to affine subspaces, i.e. planes and lines that are translated in space so that they do not pass through the origin, and to manifolds in general, i.e. curves instead of lines and surfaces instead of planes, as long as we are confined to a surrounding space of three dimensions. Then a surface, if crossed at all by the other manifold, would still have an intersection curve or point depending on whether this other manifold is a surface or a curve. If the other manifold is a point then it would typically miss the surface.

The transversality condition might be so formalized that the crossing is transverse if the sum of the dimensions of the crossing manifolds equals the sum of the dimensions of the intersection manifold and the surrounding space. If the sum of the dimensions of two manifolds is less than the dimension of the surrounding space then they miss each other because the dimension of the intersection cannot be less than zero. Hence a point and a surface in ordinary space miss each other, whereas a curve and a surface meet in isolated points. This is what we need from transversality. Poston and Stewart (1978) discuss this, as well as Morse functions.

The relation of transversality to Morse functions is as follows. The value combination of the second derivatives of a function in two arguments, for example λ_{xx}, λ_{yy}, and λ_{xy}, represents a point in three-dimensional Euclidean space. If, instead, we regard the development of these partial derivatives over time then a one-parameter manifold, i.e. a curve in space, is considered. If we had to consider two parameters then we would deal with a surface.

On the other hand, the quadratic form

$$\lambda_{xx} x^2 + 2\lambda_{xy} xy + \lambda_{yy} y^2 \tag{88}$$

which plays a crucial role in the linear differential equations that determine the behavior of the system (86–87), is degenerate in one direction if the Hessian

$$\lambda_{xx} \lambda_{yy} - \lambda_{xy}^2 \tag{89}$$

is zero without all derivatives being zero. If they are zero, the quadratic form is degenerate in two directions. If we equate (89) to zero the equation defines a surface in the three-dimensional space of which we have been talking. The surface is a double cone and contains all the points that make the quadratic form degenerate. The apex of the cones implies degeneracy in two directions. Comparing this cone of degeneracy with the manifold mentioned above, we would say that if the manifold is just a point it is unlikely to lie on this cone. If it is a curve, representing a development over time, then the cone would be intersected only at isolated points of time.

Thus, if we consider an equilibrium pattern of flows where the potential function is given, it is unlikely that the Hessian (89) is zero. If we consider a dynamic process, the Hessian could be zero at isolated points of time, but picking a moment at random we would again expect the Hessian to be nonzero. The result is that the quadratic form (88) is nondegenerate and the system of differential equations (86–87) is hence well behaved.

This implies that critical points are isolated. If we supposed the contrary, that there is some curve $x(s)$, $y(s)$ along which $\lambda_x(x,y)$ and $\lambda_y(x,y)$ are zero, then by differentiation we would obtain

$$\lambda_{xx}\,\mathrm{d}x/\mathrm{d}s + \lambda_{xy}\,\mathrm{d}y/\mathrm{d}s = 0 \tag{90}$$

$$\lambda_{yx}\,\mathrm{d}x/\mathrm{d}s + \lambda_{yy}\,\mathrm{d}y/\mathrm{d}s = 0 \tag{91}$$

However, with $\mathrm{d}x/\mathrm{d}s$, $\mathrm{d}y/\mathrm{d}s$ not both zero this system can only be solved if (89) is zero. Hence there can be no such curve, as assumed. The conclusion is that a nonzero Hessian rules out that critical points cluster along whole curves and, *a fortiori*, over whole areas. This conclusion is, however, due to transversality considerations where we have no *a priori* knowledge. Should we *know* that there is a frontier between isolated trade areas then naturally there is a curve along which the flows stagnate.

A nonzero Hessian not only rules out the possibility of critical points that are not isolated. It does the same to more complicated types of isolated critical point like the monkey saddle illustrated in Figure 2. The odd feature of a monkey saddle is that a tangent plane and the surface intersect along three different directions so that the tangent plane is divided into six sectors with their common vertex at the point of tangency such that the surface alternatingly lies above and below the plane. A nonzero Hessian, however, admits at most two directions and four corresponding sectors, i.e. an ordinary saddle.

To demonstrate this we assume $\lambda(x,y) = \text{constant}$. Differentiating twice yields

$$\lambda_{xx}(\mathrm{d}x)^2 + 2\lambda_{xy}\,\mathrm{d}x\,\mathrm{d}y + \lambda_{yy}(\mathrm{d}y)^2 = 0 \tag{92}$$

We assume also that λ_{yy} is not zero. Then (92) gives a quadratic equation in $\mathrm{d}y/\mathrm{d}x$. Should the assumption not be true but λ_{xx} be nonzero, we obtain a quadratic equation in $\mathrm{d}x/\mathrm{d}y$ instead. The roots for the two cases are, according to elementary algebra,

$$\mathrm{d}y/\mathrm{d}x = -\lambda_{xy}/\lambda_{yy} \pm (\lambda_{xy}^2 - \lambda_{xx}\lambda_{yy})^{1/2}/\lambda_{yy} \tag{93}$$

$$\mathrm{d}x/\mathrm{d}y = -\lambda_{xy}/\lambda_{xx} \pm (\lambda_{xy}^2 - \lambda_{xx}\lambda_{yy})^{1/2}/\lambda_{xx} \tag{94}$$

The roots are real only when the Hessian is negative and then there are exactly two roots, i.e. two different directions, dy/dx and dx/dy, in which $\lambda(x,y)$ is constant.

The result is reasonable. If the Hessian were positive the left-hand side of (92) would be (positive or negative) definite and the equality to zero could not hold true. Should the Hessian be zero, but not all three of the partial derivatives, then there is only one double root, yielding a whole curve of critical points. This is obvious because if the surface around a critical point has only one line of contact with the tangent plane then the contact will be that of tangency and not of intersection.

The possibility that one of λ_{xx} and λ_{yy} is zero makes no significant change to the conclusions, but only that one of the two directions of constant potential is parallel to the y or x axis. Should both partial derivatives be zero at once, the directions are parallel to both the x and y axes. Hence there are more than two directions in which the potential is constant only when $\lambda_{xx} = \lambda_{yy} = \lambda_{xy} = 0$, i.e. when we deal with a higher degeneracy. The facts are easily checked for a monkey saddle used as example. With $\lambda = x^3 - 3xy^2$ we have $\lambda_x = 3(x^2 - y^2)$ and $\lambda_y = -6xy$. The first partial derivatives are both zero only at the origin so that $x = y = 0$ is the unique critical point. However, $\lambda_{xx} = 6x$, $\lambda_{yy} = -6x$, and $\lambda_{xy} = -6y$. Hence not only the Hessian $-36(x^2 - y^2)$ but all second partial derivatives as well are zero at the critical point.

We can appeal to more powerful mathematical tools than this heuristic reasoning about what a nonzero Hessian excludes. As a result of Morse's lemma, at any critical point where the Hessian is nonzero we can introduce a smooth change of coordinates to $u(x,y)$ and $v(x,y)$ so that the potential function can be written as

$$\lambda = \pm u^2 \pm v^2 \tag{95}$$

where, for convenience, the critical point is assumed to be at the origin of (x,y,λ) space. This is a Morse saddle where the potential function is either a circular paraboloid or a hyperbolic paraboloid. The various sign combinations only result in reflections in the horizontal plane of the basic types $u^2 + v^2$ and $u^2 - v^2$.

The smooth change of coordinates can be intuitively conceived in the following way. We imagine any of these Morse saddle surfaces as marked by a continuum of vertical sticks with their lower ends fastened to a perfectly elastic rubber sheet that represents the (x,y) plane. The length of each stick represents the value of λ at the relevant point. We then deform the surface by stretching the rubber sheet in various ways, letting the sticks move with the points to which they are attached while keeping them vertical. Critical points with nonzero Hessians are then all the stationary points that we can obtain by such surface transformations. For a simple formal proof of Morse's lemma the reader may consult Poston and Stewart (1978).

It would now be tempting to deal only with potentials of the simple form (95) and study their gradient directions that yield flows that are either radial or hyperbolic in shape. This would, however, misguide us for the following reason. Even though the coordinate changes needed to arrive at a Morse saddle can be smooth there is nothing in Morse's lemma to guarantee that they can be made conformal, i.e. angle-preserving. Thus, whereas in the original coordinates the flows are orthogonal to the level curves of the potential surface this is not necessarily true in the new coordinate system. If we take an elliptic paraboloid as an example, the level curves are ellipses and the orthogonal trajectories are parabolas with a common tangent that forms the major axis of the ellipses. By a smooth and

very simple coordinate change we can transform the potential surface to a circular paraboloid. The orthogonal trajectories to the circular level curves are, of course, radials, but there exists no smooth coordinate change that maps the family of parabolas on to a family of radials, because the common tangent remains after a smooth transformation and the parabolas remain parabolas. This means that after the transformation to a Morse saddle the gradient directions give no information at all about the flow lines. In terms of the theory of differential equations we cannot smoothly transform an improper node to a proper one.

How all this is related to the subject of structural stability is understood if we recall that when we regarded a time development of the point $(\lambda_{xx}, \lambda_{yy}, \lambda_{xy})$ as a parametrized manifold in three-dimensional space, i.e. as a line, then the cone of degeneracy could be met in isolated points on the path through time. It is, however, unlikely that the parametrized manifold meets the apex (even a two-parameter manifold would not do that, so a three-parameter case is needed). Let us forget this for a moment and suppose that the apex is met at some point of time. Hence at some moment we may have a monkey saddle as illustrated in Figure 5(a), but as time passes the potential surface will be deformed. For simplicity we take the perturbation as the addition of a plane through the y axis. This means that a slight component of horizontal flow is added to the flow lines of the monkey saddle. Formally we deal with the potential $\lambda = x^3 - 3xy^2 - xt$, where t denotes time. The monkey saddle is relevant just for one point of time, $t = 0$. For any time before, however close to zero, the pattern looks like that in Figure 5(b), and for any time afterward, however close to zero, the pattern looks like that in Figure 5(c). This illustrates the instability of the monkey saddle flow.

As the monkey saddle point in Figures 5(b) and (c) is split into two ordinary saddles this will contrast with what will be said below concerning the improbability of a trajectory joining two saddle points. This hints at the fact that the cases portrayed are still not stable and that they may be split by additional perturbations. We can see this by perturbing the flow illustrated in Figure 5(c), adding a slight vertical flow to the horizontal one already introduced. Putting $\lambda = x^3 - 3xy^2 - xt - yt$ produces the pattern shown in Figure 5(d) for positive t, no matter how close to zero. Even though the two saddles remain there is no longer a trajectory joining the saddles, as there was in Figure 5(c). This illustrates that a saddle connection is still structurally unstable and can be broken by the slightest further perturbation.

We have by no means detected all the phenomena that can arise from perturbing a monkey saddle. We have introduced a sample of two perturbations but, as was said in the introduction to this discussion, the case represents a degeneracy in two directions that is likely to occur only with a three-parameter family of perturbations. There is even evidence that the universal unfolding of a gradient flow is of higher dimension than the universal unfolding of its potential function. With two-dimensional flows one more parameter is added. The final structurally stable configuration into which a monkey saddle can be split seems to be a set of four critical points: one sink, one source, and two saddles of the ordinary type.

Figures 5(e) and (f) illustrate the results of perturbation of a monkey saddle by a radial flow. The potential is $\lambda = x^3 - xy^2 - x^2t - y^2t$. Figure 5(e) illustrates the situation just before time zero and Figure 5(f) the situation just afterward. For time zero, of course, Figure 5(a) portrays the situation, but the situation can still be altered by further perturbations. We have only illustrated a small sample of the phenomena that can arise with the

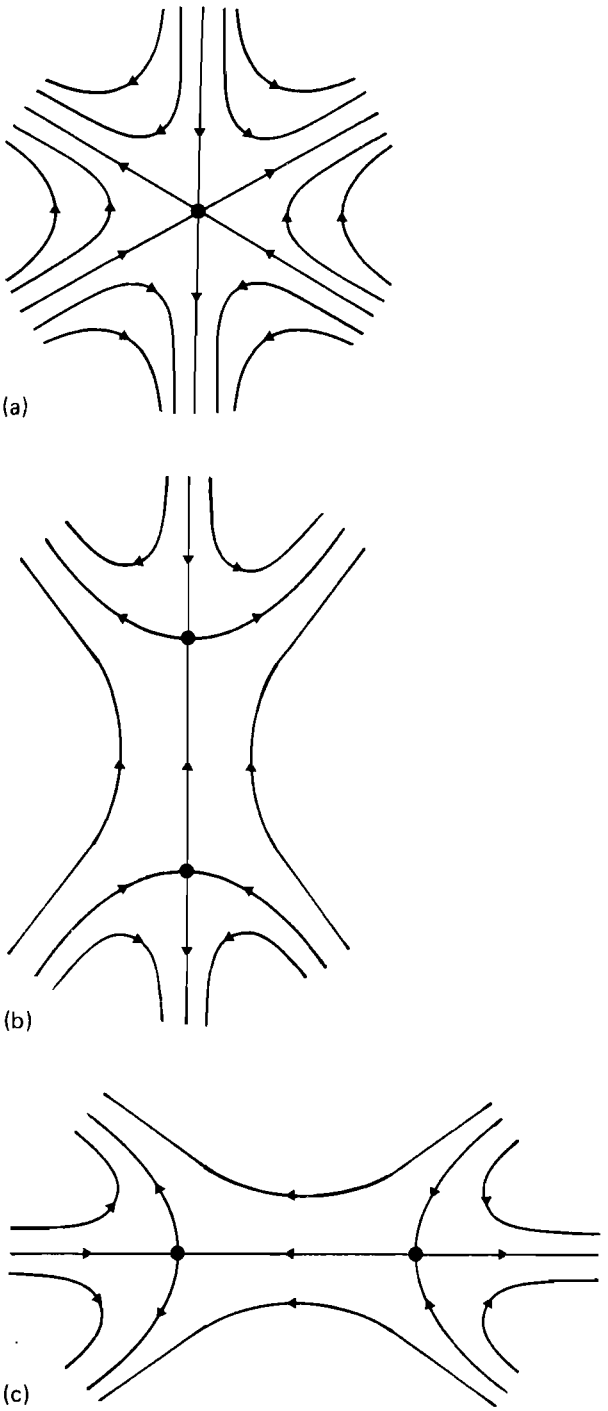
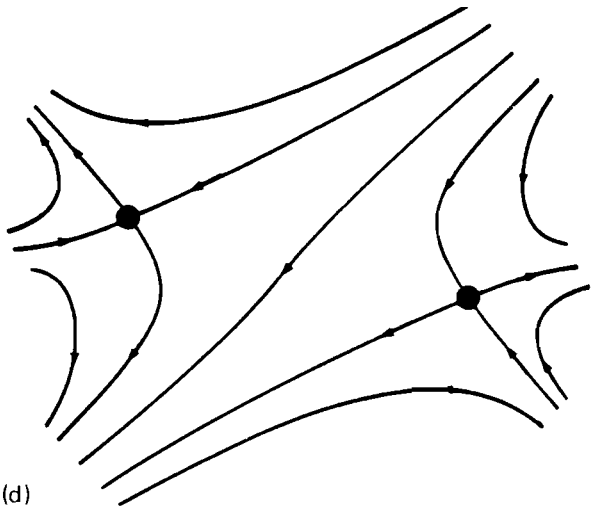
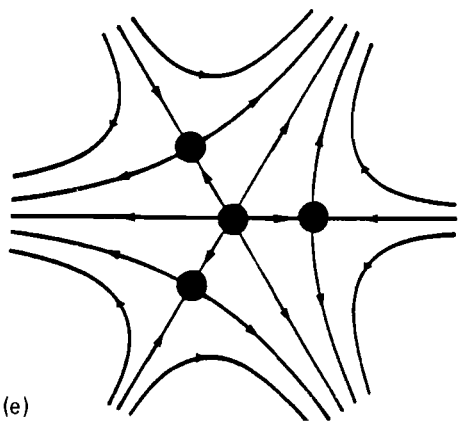


FIGURE 5 Monkey saddle (a) at an isolated moment, (b) just before, (c) just afterward.





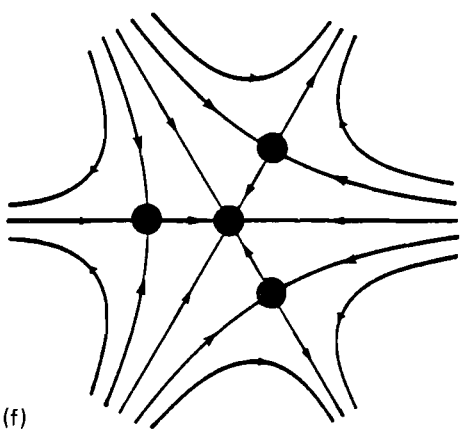


FIGURE 5 (d) Monkey saddle perturbed by slight horizontal and vertical flows. (e) Splitting the monkey saddle by weak radial outward flow, (f) by weak radial inward flow.

extremely unstable monkey saddle. This is of some importance because flow patterns of the monkey saddle type are associated with hexagonal subdivisions of a region, as is traditional in the Christaller–Lösch paradigm. We shall return to this issue.

13 NODES AND SADDLES

Even though it is impossible to proceed from potentials transformed to Morse saddles, we know a lot about the critical points already. From analyticity of the potential we know that the critical points are simple and by the transversality considerations we know that they are isolated. Hence, the solutions to (86–87) are of the same type as those of a linear system,

$$\dot{x} = \lambda_{xx} x + \lambda_{xy} y \tag{96}$$

$$\dot{y} = \lambda_{yx} x + \lambda_{yy} y \tag{97}$$

with a nonzero Jacobian, i.e. $\lambda_{xx}\lambda_{yy} - \lambda_{xy}^2 \neq 0$ (Simmons 1972). The categories are nodes (proper and improper), saddles, and spirals (including centers). Among these we can even exclude spirals and centers. This is so because the differential equations are obtained from a potential function. The roots of the characteristic equation of the system (96–97) are

$$\kappa_1, \kappa_2 = (\lambda_{xx} + \lambda_{yy})/2 \pm [(\lambda_{xx} - \lambda_{yy})^2 + 4\lambda_{xy}^2]^{1/2}/2 \tag{98}$$

The root structure determines the character of the critical point. Complex roots lead to spirals and purely imaginary roots to centers. If the roots are real, as they obviously are in our case, then we deal with nodes or saddles, depending on whether the roots have the same or opposite signs. The signs are the same when the Hessian is positive and opposite when it is negative. Spirals were ruled out since $\lambda_{xy} = \lambda_{yx}$ with a potential flow, and other types were ruled out because the potential was assumed to be analytic and because of our appeal to transversality.

Our assumptions have been hardly restrictive at all. The flow lines could be obtained from a potential because this was the consequence of the solution of the optimum path problem. Analyticity is a very general property and transverse crossings deal with everything that is stable and robust to disturbances in a system. The conclusions are quite like those conjectured by the (incorrect) reasoning about the gradient directions to Morse saddles. The only difference is that with a Morse saddle any node would be proper, whereas proper nodes at present only occur if $\lambda_{xx} = \lambda_{yy}$ and $\lambda_{xy} = 0$.

In the discussion about how different critical points are related to each other in the region considered, we shall find that the improper nodes turn out to be typical, rather than the proper ones. However, the relation of a proper node with its radial paths and circular price contours to von Thünen's model makes the case particularly interesting. In addition, we are going to study a regular saddle, where movement is on hyperbolas and where the surroundings of the critical point are split into sectors of different kinds.

14 LOCATION PATTERNS AROUND NODES AND SADDLES

It is now time to discuss simple location patterns that can arise around a critical point of the node or saddle type. We shall deal with proper nodes. Let us assume a potential of the form

$$\lambda = (x^2 + y^2)^{1/2} \tag{99}$$

This is simply the square root of the basic Morse case of a circular paraboloid. We have taken the square root because its gradient must fulfill the conditions of (26) and (27) for a convenient f:

$$\nabla\lambda = (x/\rho, y/\rho) \tag{100}$$

From (26) and (27), with $f = 1$,

$$\boldsymbol{q} = |\boldsymbol{q}|(x/\rho, y/\rho) \tag{101}$$

$$\boldsymbol{l} = -|\boldsymbol{l}|(x/\rho, y/\rho) \tag{102}$$

The simplest assumption to make is that we study some disk-shaped neighborhood of the critical point and that, as a consequence, the norms of the flows are functions with circular symmetry. As $|\boldsymbol{q}|$ and $|\boldsymbol{l}|$ are then functions of $\rho = (x^2 + y^2)^{1/2}$ only, we obtain simple expressions for the divergences of (101) and (102), namely

$$\nabla . \boldsymbol{q} = \frac{1}{\rho}\frac{\mathrm{d}}{\mathrm{d}\rho}(|\boldsymbol{q}|\rho) \tag{103}$$

$$\nabla . \boldsymbol{l} = -\frac{1}{\rho}\frac{\mathrm{d}}{\mathrm{d}\rho}(|\boldsymbol{l}|\rho) \tag{104}$$

Hence we conclude that the critical point is the center of some industrial region if the divergence of the flow of goods is positive, i.e. $\nabla . \boldsymbol{q} = q - q' > 0$, whereas the divergence of the flow of labor is negative, i.e. $\nabla . \boldsymbol{l} = l' - l < 0$. The condition for this to hold is that the flow intensities of goods as well as labor should decrease as we approach the critical point, or if they increase the rate of increase should be lower than the rate of decrease of the distance to the center. Intuitively, the assumption seems reasonable as the flows accumulate through the surrounding industrial region. The conclusions would be reversed in a residential region because all the signs would be reversed, starting with that of the potential (101).

Let us now study the industrial center a little more closely. Use of (101) in (30) and (31) yields the conclusion that prices increase with the distance from the center, whereas wages decrease. Both are constant on concentric circles, as is the real wage rate that decreases with the distance from the center. This is acceptable to intuition as the residents of a certain location have the option to work closer to their homes for a lower wage or to receive a higher wage closer to the center and have to deduct commuting costs. Likewise,

they could buy goods closer to their homes at higher prices or buy them closer to the site of production and add transportation costs. Equilibrium requires that these options are equivalent.

The varying real wage entails different choices of technique at the various sites of production. According to (8) and (9) a labor-intensive technique of production would be chosen at the outskirts of the region because of low real wages, whereas capital-intensive production occurs close to the industrial center owing to high real wages. We can also see that according to the same equations factor productivities are high for both capital and labor in the industrial center. If we do not interpret the homogeneous product assumption literally we could say that basic production of goods takes place in the center, whereas production of services mainly takes place in the outskirts of the industrial region.

Some more conclusions can be drawn. From (8) and (10) we see that, because capital rent is constant, the land rent g is proportional to the density of capital, k, invested per unit land area. It is now reasonable to suppose that capital is concentrated and land rent is high close to the industrial center, but then we see from (10) that because land rent is high and prices are low in the center, production per unit land area, q, is high there. All this is appealing to intuition.

We can also say that the fact that land rent is high in the center could explain why workers do not live only where real wages are most advantageous. As land rent determines housing costs and the real wage variation represents communication costs we have the disadvantage of expensive housing at the center along with the advantage of low transportation costs, whereas the matters are reversed at the outskirts. This is reasonable for an equilibrium pattern of residential location. The case is illustrated in Figure 6.

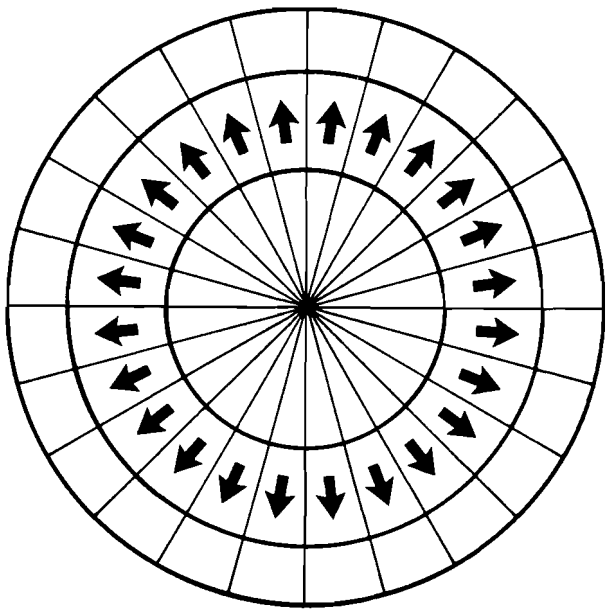


FIGURE 6 Flow and potential contours at a proper node.

The second interesting case is

$$\lambda = (x^2 - y^2)/2 \tag{105}$$

representing the standard Morse case of a hyperbolic paraboloid. We have divided by two because then the gradient is

$$\nabla\lambda = (x, -y) \tag{106}$$

The length is $\rho = (x^2 + y^2)^{1/2}$, so for $f = \rho$ the potential is in accordance with (26) and (27). From these equations,

$$\boldsymbol{q} = |\boldsymbol{q}|(x/\rho, -y/\rho) \tag{107}$$

$$\boldsymbol{l} = -|\boldsymbol{l}|(x/\rho, -y/\rho) \tag{108}$$

We can now make various assumptions about the Euclidean norms of the flows. One possibility is to assume circular symmetry again, so that $|\boldsymbol{q}|$ and $|\boldsymbol{l}|$ are functions of $\rho = (x^2 + y^2)^{1/2}$ only. Then we have from (107) and (108) the divergences

$$\nabla_{\circ}\boldsymbol{q} = \frac{x^2 - y^2}{\rho} \frac{\mathrm{d}}{\mathrm{d}\rho} \frac{|\boldsymbol{q}|}{\rho} \tag{109}$$

$$\nabla_{\circ}\boldsymbol{l} = -\frac{x^2 - y^2}{\rho} \frac{\mathrm{d}}{\mathrm{d}\rho} \frac{|\boldsymbol{l}|}{\rho} \tag{110}$$

Provided that the ratios of the flow intensities to the distance from the origin are monotonic functions of this distance, everything about excess supplies and demands is decided by the sign of $x^2 - y^2$. From (107) and (108) we see that goods and labor flow along hyperbolic paths in four quadrants, labor from east and west to north and south and goods in the reverse directions. It is reasonable to assume that then there is excess demand of labor and excess supply of goods in north and south, whereas there is excess supply of labor and excess demand of goods in east and west. Accordingly the derivatives in (109) and (110) should be negative, so that if the flow intensities increase with the distance from the origin they do so at a lower rate than the distance itself. In that case the model is consistent.

The assumption of circular symmetry is not at all as natural in the present case as in the case of a proper node. Even if we specify the region as circular there may be reasons to regard other structures. The assumption may, however, serve as an illustration because it is in no way unreasonable. As the flow lines are hyperbolas they come closest to the origin on the lines at $\pm 45°$. If the flows are built up by additional elements on one side of the lines whereas elements are only withdrawn from the other side it seems acceptable that the flows have maximum force in their middle sections, where they are close to the origin.

The resulting structure has two industrial sectors, one north and one south, and two residential sectors, one east and one west. We have seen that hyperbola-shaped flows are orthogonal trajectories to another family of hyperbolas rotated by an angle of 45°. Hence, as can be seen from (105) in conjunction with (30) and (31), prices increase from east and west to north and south, whereas wages do the reverse. *A fortiori*, real wages are high in the north and south and low in the east and west. This means that the sectors in the north and south are industrial, having excess demand of labor and excess supply of goods, whereas the facts are reversed for the sectors in the east and west, which hence have residential character. Goods flow from the industry to the residences and labor flows in the reverse direction. Along the flows the local price and wage increase so that in the industrial sectors we encounter use of a capital-intensive technique, whereas labor-intensive production occurs in the residential sectors. Again assuming much capital to be invested per unit area in the industrial sectors, we find high land rents there along with a high concentration of productive activity.

The case is illustrated in Figure 7. Hence the two location patterns outlined in Figures 6 and 7 are the typical organizations around a critical point of generic type, i.e. one of a stable flow. This, of course, also holds if all the flows are reversed, by changing the sign of the potential. The saddle case does not change character, but the node becomes one where there is a reservoir of labor and consumptive potential at the origin, whereas production occurs in the outskirts. For agricultural production this is a von Thünen case.

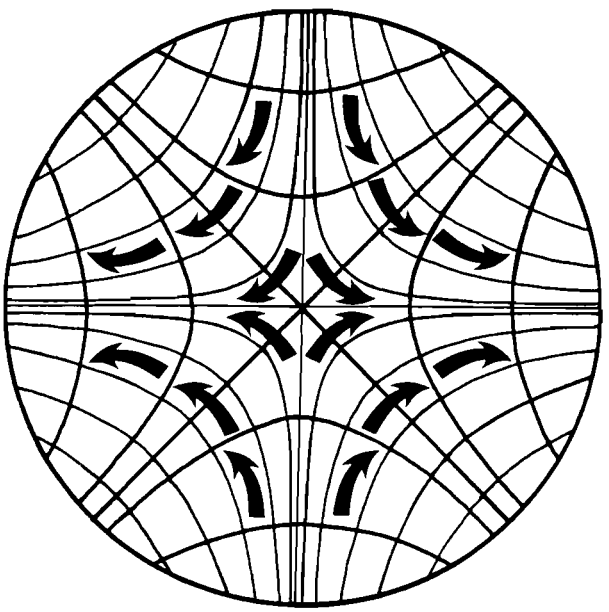


FIGURE 7 Flow and potential contours at a saddle.

Before looking at how the whole picture could be assembled to yield a typical flow pattern and a corresponding location pattern, we have to return to the question of structural stability and find how different critical points can be related.

15 TRAJECTORIES CANNOT JOIN SADDLES

The fundamental tool that we use in assembling the whole picture is the fact that it is extremely improbable that two saddles are directly joined by a trajectory. There are only two different directions around a saddle in which the saddle point itself is connected with the system of trajectories. Of the infinity of trajectories in the neighborhood of the critical point there is one pair going out from it in opposite directions and one pair going into it from opposite directions. Each pair represents a characteristic direction, and these two directions correspond to the two real roots (of opposite signs for a saddle) of the characteristic equation. Hence, only the two characteristic lines pass through the critical point whereas all the other trajectories miss it, following hyperbolic paths in one of the four quadrants into which the characteristic lines split the plane around the critical point. As two saddles are joined only if they have a critical line in common this would seem highly improbable.

As we deal with potential flows the system of differential equations corresponds to simple geometric properties of the potential surface $\lambda(x,y)$ in (x,y,λ) space. The mapping of (x,y) on to $(x,y,\lambda(x,y))$ is a well defined parametrization of a surface in ordinary space. The tangent vectors to the coordinate lines are $(1,0,\lambda_x)$ and $(0,1,\lambda_y)$. Accordingly, the Gaussian "first fundamental coefficients" that determine the metric structure of a surface are obtained as various dot products of these tangent vectors. In standard notation,

$$E = (1,0,\lambda_x) \cdot (1,0,\lambda_x) = 1 + \lambda_x^2 \tag{111}$$

$$F = (1,0,\lambda_x) \cdot (0,1,\lambda_y) = \lambda_x \lambda_y \tag{112}$$

$$G = (0,1,\lambda_y) \cdot (0,1,\lambda_y) = 1 + \lambda_y^2 \tag{113}$$

The "second fundamental coefficients" that define the whole curvature structure of the surface will also be needed. To define them we need the second derivatives of the space vector $(x,y,\lambda(x,y))$ with respect to the coordinates (x,y). These are $(0,0,\lambda_{xx})$, $(0,0,\lambda_{xy})$, and $(0,0,\lambda_{yy})$. The unit normal vector to the surface, $(-\lambda_x, -\lambda_y, 1)/(1 + \lambda_x^2 + \lambda_y^2)^{1/2}$, is also required. The three second fundamental coefficients are defined as dot products of the normal vector with each of the second-order derivative vectors listed. Thus

$$L = \lambda_{xx}/(1 + \lambda_x^2 + \lambda_y^2)^{1/2} \tag{114}$$

$$M = \lambda_{xy}/(1 + \lambda_x^2 + \lambda_y^2)^{1/2} \tag{115}$$

$$N = \lambda_{yy}/(1 + \lambda_x^2 + \lambda_y^2)^{1/2} \tag{116}$$

The curvature in the direction of the normal of the surface taken in a section of the direction dy/dx is given by

$$\kappa = \frac{L(dx)^2 + 2M\,dx\,dy + N(dy)^2}{E(dx)^2 + 2F\,dx\,dy + G(dy)^2} \tag{117}$$

In Gaussian differential geometry there are two directions in which curvatures are maximal and minimal, respectively. In a hyperbolic point, i.e. at a saddle, they are of opposite sign. These directions are given by the system

$$(L - \kappa E)\,dx + (M - \kappa F)\,dy = 0 \tag{118}$$

$$(M - \kappa F)\,dx + (N - \kappa G)\,dy = 0 \tag{119}$$

which only has a solution when the determinant of the system is zero, i.e. when

$$(L - \kappa E)(N - \kappa G) - (M - \kappa F)^2 = 0 \tag{120}$$

If we study the system of relations introduced at a critical point with $\lambda_x = \lambda_y = 0$, then $E = G = 1$ and $F = 0$ from (111–113), which means that the surface is locally isometric to the plane. Moreover, from (114–116), $L = \lambda_{xx}$, $M = \lambda_{xy}$, and $N = \lambda_{yy}$, which gives (120) the form

$$\kappa^2 - (\lambda_{xx} + \lambda_{yy})\kappa + (\lambda_{xx}\lambda_{yy} - \lambda_{xy}^2) = 0 \tag{121}$$

This second-order equation has the solutions

$$\kappa_1, \kappa_2 = (\lambda_{xx} + \lambda_{yy})/2 \pm [(\lambda_{xx} - \lambda_{yy})^2 + 4\lambda_{xy}^2]^{1/2}/2 \tag{122}$$

which, when substituted into (118–119), yield two principal directions $(dy/dx)_1$ and $(dy/dx)_2$. If we start from the critical point in one of these principal directions and continue in the direction of the gradient, we can trace a curve on the potential surface. Since there are four (oriented) directions of this kind at a critical point we can trace four such curves. How each of these curves lies on the surface depends on the global character of λ. By introducing a suitable deformation of the surface, we can make any curve starting out in a principal direction from one critical point a gradient curve while keeping this point and any other critical point unaffected.

Now let us consider another critical point on the surface. At present, we are dealing with a one-dimensional curve and a zero-dimensional point embedded in a two-dimensional surface. Because of transversality they typically miss each other. The reader is referred to do Carmo (1976) for the terminology of classical differential geometry.

Let us now compare eqns. (98) and (122), which yield the roots of the characteristic equation for the dynamic system and the principal curvatures of the potential surface, respectively. They are identical and hence the projections of the principal directions on to the parameter plane are the same as the characteristic directions. This suggests that the projections of the surface curves are simply the trajectories running through the saddle, but then it is extremely unlikely that any one of these trajectories runs through another saddle point. The reader may ask why it would not be equally unlikely that a node lies on one of these trajectories. The answer is given if we reverse the roles and regard the trajectories running out from a node. On the potential surface they represent a collection of radiating curves going in all the gradient directions. It is in no way unlikely that another critical point, a node or a saddle, lies on one of these infinitely many trajectories.

16 THE LATTICE OF CRITICAL POINTS

With the conclusion of the preceding section that no trajectory can join two saddles, let us see how a lattice of critical points can be put together. Starting with a saddle, we know that it is intersected by two trajectories. Let us suppose that, as we follow these two trajectories to both sides of the critical point, sooner or later we meet another critical point. This is the most complicated possible arrangement, but since no trajectory can lead from one saddle to another we know that all four new critical points met are nodes. We can draw a picture on a regular quadratic point lattice as the two principal directions give a quadratic structure to the arrangement. If we hence mark one point as a saddle in a quadratic point lattice then it will be surrounded to the east, north, west, and south by neighboring nodes. In the lattice we can connect the points by sets of horizontal and vertical straight lines. As two directions are inward to a saddle and two outward from it one pair of surrounding nodes will be an attractor or a sink and the other pair a rejector or a source.

We arrange the critical points in a quadratic lattice because this represents the most general pattern for arranging them, but from this basic shape we can transform the system of critical points and flows to any geometric shape wished. Hence the pattern covers a great many patterns if we take smooth coordinate transformations into account, but it should be observed that however we transform the quadratic grid with its lattice of intersection points the basic shapes will still be quadratic rather than hexagonal. This holds true for various subregions, trade areas, and so on. Hence structural stability seems to contradict the economic principle of packing market areas in a honeycomb arrangement, and hence the Christaller–Lösch theory of market areas (Beckmann 1968).

We can now trace out the whole typical flow. We use the quadratic grid of Figure 8 to represent trajectories, whose directions are marked by arrows. We have already inferred

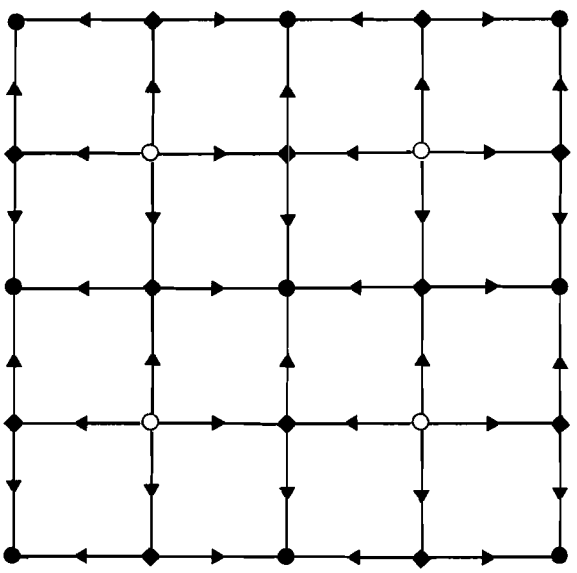
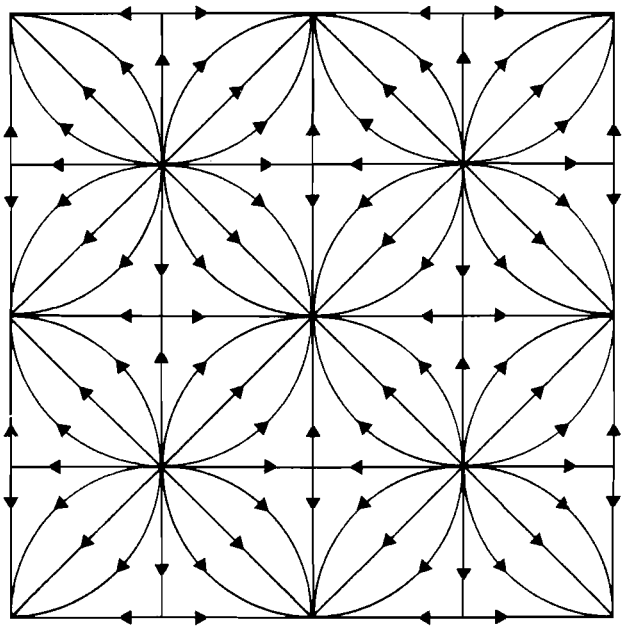
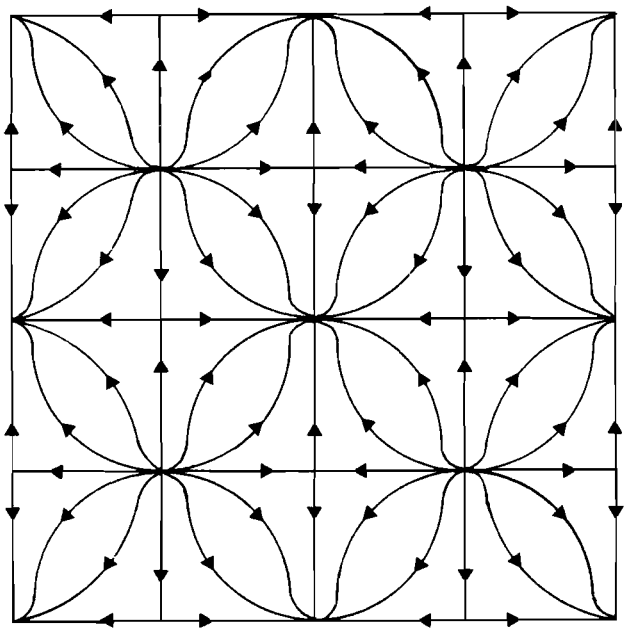


FIGURE 8 Oriented graph of the basic structurally stable flow. Critical points: ♦ saddles, ○ sources, • sinks.



(a)



(b)

FIGURE 9 (a) Flow lines that are Minkowski circles. (b) The basic structurally stable flow.

that a saddle (♦) has two sources (○) in two directions and two sinks (●) in the orthogonal directions as its closest neighbors. This defines the directions of the trajectories incident to the saddle. However, since the characters of the four surrounding nodes are given, all lines incident to them are given certain orientations. These lines meet two by two at right angles. We conclude that at these points of intersection there are ingoing as well as outgoing paths. Hence they are saddles.

Knowing that saddles are surrounded by nodes, we can proceed farther away from the original saddle, determining by the same argument the character of each point and the orientation of each edge of the graph. There is hence one basic lattice arrangement and a basic oriented graph structure of the flows. It is shown in Figure 8. Of course, there are lacking whole families of trajectories that would represent whole flow patterns. We shall supply such an example below, but it should be noted that the graph structure fully adequately represents the basic flow. It would be very simple to fill in families of directed trajectories as soon as the graph frame is given.

Figure 9(a) shows an example that has a mathematical representation:

$$|x - m|^{\mu} + |y - n|^{\mu} = 1 \tag{123}$$

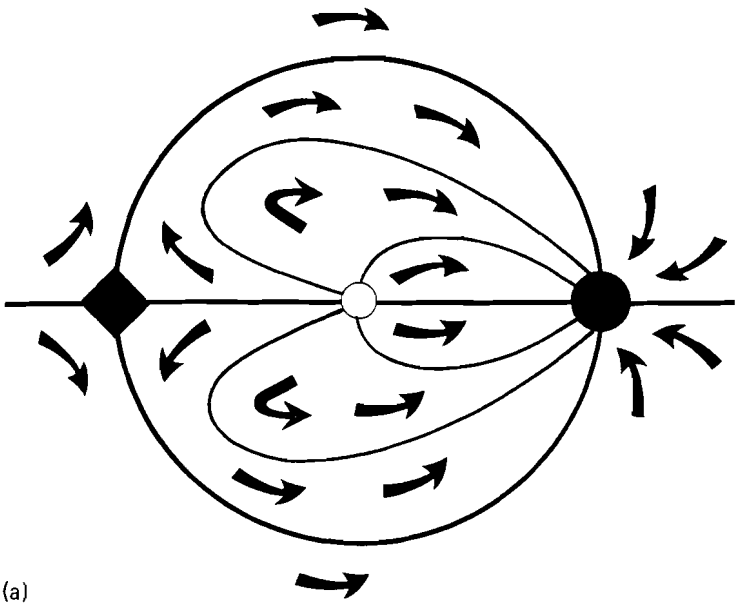
where m, n are integers such that $m + n$ is even and where μ varies between zero and unity. The relation determines a set of concentric Minkowski circles for various μ when m, n are fixed. Changing the latter generates a new set of Minkowski circles so that the whole lattice of squares is filled by lines. Different flows can be obtained from the pattern in Figure 9(a) by smooth coordinate changes.

Strictly speaking, the case portrayed is still not structurally stable since all the nodes are proper (or foci in another terminology), which means that the eigenvalues are equal. It is easy to perturb a differential equation so that only one eigenvalue is changed and hence the case is not stable as far as the shape of the trajectories is concerned. A stable flow would have to look something like Figure 9(b). Obviously this gives a main direction to each node as all trajectories except one pass it in that direction.

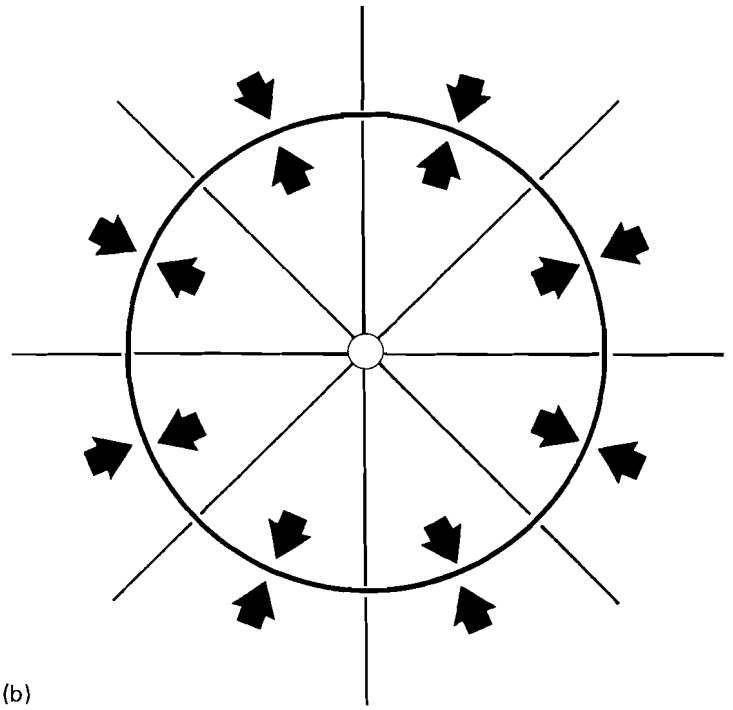
17 POSSIBLE COMPLICATIONS OF THE BASIC PATTERN

We assumed quite arbitrarily that all saddle trajectories, i.e. the four trajectories actually incident to the saddle point, should be incident to *different* nodes. We saw that saddles could not be connected by trajectories, which implied that an outgoing trajectory did not return to the same saddle after a loop as an ingoing trajectory. This does not exclude the possibility that a pair of trajectories, outgoing or ingoing, can be incident to the same node. This cannot be the case with both pairs since the trajectories would have to cross, but for one pair this is perfectly possible. That pair would form a closed loop, with a node between, as shown in Figure 10(a). As there is a trajectory from the saddle going inside this loop that cannot end at the opposite node, because it has a wrong direction, we conclude that there must be a node inside this loop. Thus we arrive at the organization illustrated.

In particular, it should be observed that the circular loop defines an *isolated trade area*. Inside the circle everything consumed is produced inside as well. This isolated trade



(a)



(b)

FIGURE 10 Flows and trade areas: (a) structural stability, (b) structural instability.

area is different to the one usually encountered in location theory, which usually looks somewhat like Figure 10(b), where the flow stagnates along the boundary of the area. Exactly because of this, the traditional picture is structurally unstable. As we have seen, singular points in a structurally stable flow are isolated and do not cluster along lines.

However, the case of a boundary consisting of stagnation points has a strong intuitive appeal, so we should say something about it. The classical case is the market area separation in Launhardt's theory. If the supplier with higher production costs has lower freight rates then the trade areas are inside and outside an elliptic boundary curve. The case is then like that of Figure 10(b), except for the shape of the boundary.

Let us suppose now that transportation facilities are not uniform and that there are particularly good transportation facilities along the main connection between the two supply centers. Then the routes are not radial, but would be curved as if they were attracted by the particularly favorable transportation facilities. The case then turns into that depicted in Figure 10(a). The isolated areas remain, but the flows do not actually stagnate on the boundary. Rather, there is a flow, however weak, along it. A moment of reflection will show that what we really want to keep from location theory is the concept of trade areas, rather than the conclusion that there must be stagnation everywhere along the boundary.

The complications brought into the basic pattern, Figure 8, are not as large as could be imagined. This may be most easily seen if we consider the "price landscape," or the picture of the potential surface along which the flows may be traced in the gradient directions. This price landscape is shown in Figure 11, which corresponds to the flow of Figure 8. Price maxima, at consumption centers, alternate with price minima, at production centers, and in between there are saddle points. The only thing we have to do to bring the present considerations into the picture is to introduce "craters" on the tops of the hills

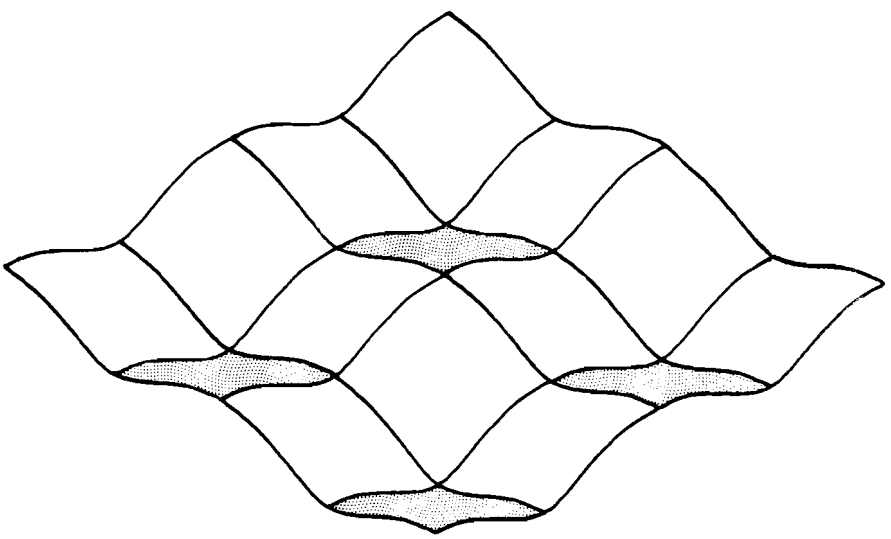
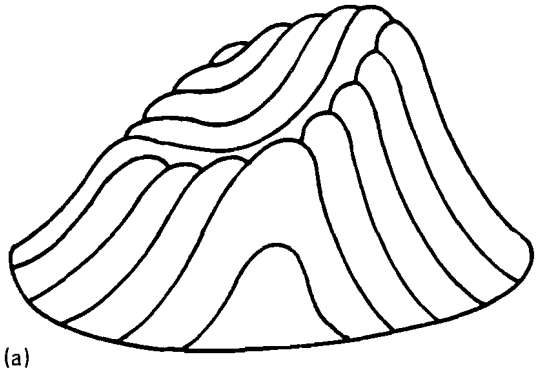


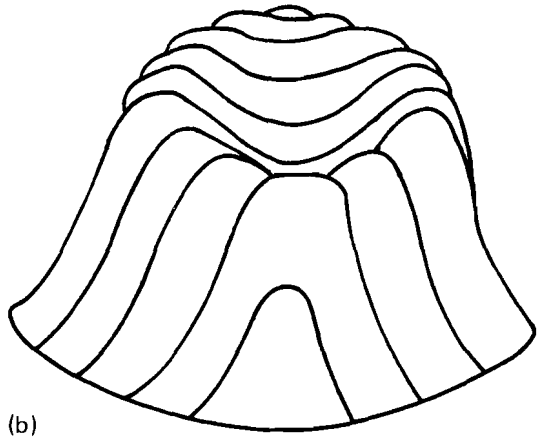
FIGURE 11 Price landscape for the basic quadratic flow.

(and, *mutatis mutandis*, for the bottoms), as shown in Figure 12(a). We only have to remember that the craters cannot look as in Figure 12(b): this would lead to the flow of Figure 10(b), and hence to instability. Figure 12(a) corresponds to the flow of Figure 10(a). These deformations of the landscape of Figure 11 can be made at any set of nodes. By considering these we actually cover everything that is compatible with structural stability.

The change is interesting from another point of view, too. Obviously, the bottoms of the craters need not be as deep as the bottoms of the original landscape. We can hence introduce a notion of hierarchy of supply centers. As the procedure just discussed can be repeated at the new nodes of the system we can actually nest trade areas inside each other and so consider a hierarchy of any number of levels.



(a)



(b)

FIGURE 12 Price landscapes and trade areas: (a) structural stability, (b) structural instability.

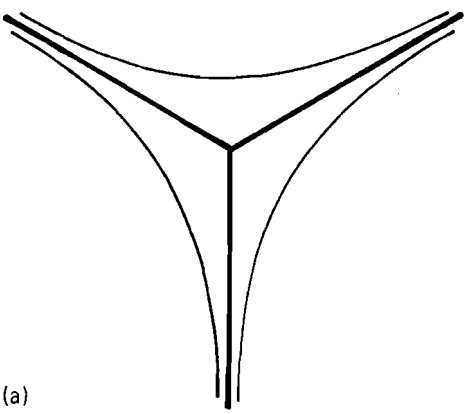
18 INCOMPATIBILITY OF HEXAGONAL SHAPES WITH STRUCTURAL STABILITY

If we tried to distinguish a pattern of various subregions in Figures 8 and 9 we would end up with a chessboard of alternating industrial and residential areas. The squares of the chessboard would have a side length inflated by a factor of $2^{1/2}$ compared with the side of the quadrats in the flow pattern and would be tilted at 45° to the latter. The flow and the subdivision patterns would fit together so that each square of the subdivision would have each one of its corners in a saddle point and have a node as its center.

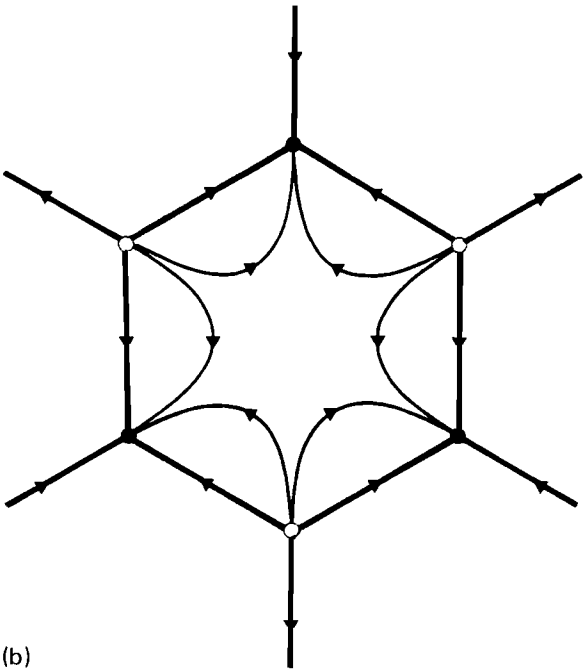
The basic spatial arrangement is thus a quadratic regular tessellation. As already mentioned, this is not in agreement with the hexagonal shapes that have been basic to regional modeling since Christaller's work. Due to the mathematical properties of the hexagonal tessellation, i.e. that it encloses the largest area for a given total perimeter among all the regular tessellations, it comes as close to the isoperimetric solution of a circle as is possible with densely packed cells. For the mathematics of tessellations we refer to Fejes Tóth (1964). Tessellations occur also in studies of the economic packing of market areas (Beckmann 1968).

These shapes, however, would be connected with flows of the same type. We would expect critical points to which there are three incident directions at 120° angles, or six incident directions at 60° angles. The former are in fact completely inorientable. If we try to draw trajectories in the three 120° angles and to give them an orientation we inescapably arrive at a contradiction. This proves that we cannot have anything of "saddle"-like character (Figure 13(a)). On the other hand, such a junction of three incident directions could well be a node. However, if we put together a set of only nodes, as in Figure 13(b), we produce a hexagon with alternating sources and sinks at the corners. Completing the picture of flow lines inside this hexagon, however, shows that there is something missing in the middle of it, which can only be a monkey saddle. Thus we are left with the case of six incident directions. They too can be made nodes, as in Figure 13(c), but then they must have monkey saddles on the surrounding hexagon.

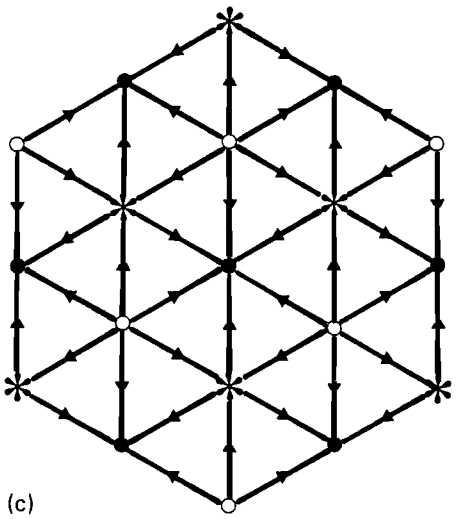
We conclude that if we try to arrange the graph of a flow on a hexagonal pattern, then provided that we at least construct a consistent pattern we end up with monkey



(a)



(b)



(c)

FIGURE 13 (a) Inorientable flow. (b) Missing monkey saddle. (c) Occurrence of monkey saddles (*) in hexagonal lattice.

saddles, which are structurally unstable. The case is not strong enough to dismiss hexagonal patterns in favor of quadratic ones, but there is at least a hint that the optimum hexagonal patterns may be structurally unstable.

We have been talking of flow patterns, but the resulting region shapes are similar to the flow patterns. This is illustrated in Figure 14, where the basic subdivision of the region can be interpreted as either triangular or hexagonal. This way of fitting flows to these shapes seems to be the only, unique possibility.

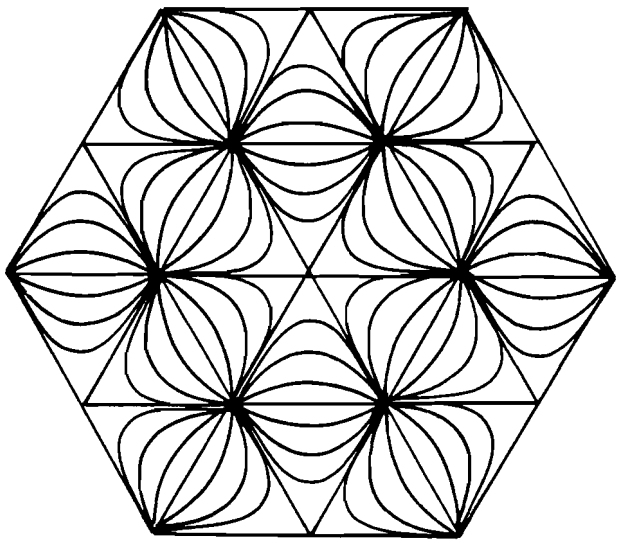


FIGURE 14 Flows associated with hexagonal or triangular subdivisions.

The considerations presented above have an uncommon character; at the same time, they contradict some deeply rooted ideas concerning spatioeconomic structure. Therefore, they should be discussed in some more detail at the intuitive level. Classical location theory, associated with the names of Launhardt–Weber and Christaller–Lösch, has a geometric character in the Euclidean sense. We find boundaries of market areas to take the shape of conic sections, and find the market areas of identical firms in homogeneous space to be packed as regular tessellations of hexagons.

In contrast to this geometric information of great Euclidean detail, our characterization is in terms of topological "rubber-sheet geometry." It may seem surprising that this vague information is at all able to refute anything as precise as the beautiful hexagonal tessellations of classical theory, but so it is, and we have to resolve the contradictions.

First, it should be stressed that our considerations imply nothing at all about the original works of the classical authors, all of whom treat the setting of a homogeneous plane with equal facility of communication everywhere. This spatial invariance of transportation costs and the resulting straight-line routes of communication make all the classical models linear. Linear models automatically possess structural stability. Only when we wish to derive the results due to the classics from nonlinear models do the difficulties arise. Structural stability is then no longer guaranteed, but must be expressly assumed. As

some basic features of classical location theory are refuted, we have to make a choice. We could abstain from dealing with nonlinearities and stick to the classical models of the homogeneous plane, or we could insist on nonlinear models. Then there is a conflict. We can dismiss structural stability as irrelevant, or else we have to dismiss traditional market areas and hexagonal tessellations.

Spatial economists, like Isard (1956), have long been aware of the severe limitations imposed by linearity. Theoretical geographers have used their understanding of map projections to define curved spaces, in which efficient communication is along as straight lines as exist on a curved space (geodesics) (e.g. Warndtz 1967, Angel and Hyman 1976). The obvious purpose has been to make it possible to apply classical location theory to the more general case of inhomogeneous space. We have concluded that certain features of classical theory do not easily carry over to the more general case. It is unfortunate that some of these features are among the most appealing in classical theory.

What about not assuming structural stability? Unfortunately, we would need a non-scientific amount of self-confidence to claim that any model we construct in the social sciences takes explicit care of all relevant factors and interactions so that we can disregard exogenous disturbances. Hence, we cannot escape admitting that our model is subject to perturbation. This being so, it is a mathematical fact that it is extremely unlikely that we shall observe a structurally unstable flow. Another mathematical fact is that accumulated singularities and monkey saddle singularities are vanishingly unlikely to be observed.

Just how bad is this? For one thing, we no longer have trade areas where trade stagnates along the boundaries, but we still have completely self-contained trade areas, delimited by boundaries that are crossed by no trajectory. The difference is that now the boundaries themselves are made up of sections of trajectories. With reference to the earlier discussion, this does not seem too revolutionary. We saw that nonlinearity could develop in a traditional Launhardt–Weber model if a transportation network was developed along the axis of connection between the sites of the firms. Efficient routes would then be curved because of the attraction toward this axis. This curving of flow lines makes them asymptotic rather than transverse to the boundary itself. From this there is only a small step to our new concept of a market area.

The harm done to the hexagonal tessellations of economic space is greater. To be more exact, if *all* trade is confined to separate market areas (in one single subdivision) then the concern is still with the boundaries of market areas and the accumulation of singularities only. If we recall how Lösch organizes space by several superposed hexagonal tessellations of increasing cell size, we find that flows of goods of some levels of the hierarchy actually cross boundaries to other levels. The Löschian communications are still along straight lines, but once we make the model nonlinear, and still wish to maintain the hexagonal patterns, we have to realize that this cannot be done without monkey saddles and that the monkey saddles are structurally very unstable. Thus, the Christaller–Lösch theory cannot be applied to inhomogeneous space.

19 CATASTROPHES

We have used structural stability as a basic modeling instrument. The idea has been that only structurally stable flows and location patterns are likely to persist through time. The structurally unstable configurations were discarded because, in the course of their

development, they would only have momentary existence. The structure before and after those isolated moments would have an easily recognizable, qualitative character and would only change smoothly over time. This character would be one of structural stability.

The story is, however, not as simple as that. Even if the structurally unstable pattern is itself without interest, owing to its only momentary existence, the fact that the stable patterns before and after are qualitatively very different is of great importance. We can take Figure 5 as an example. The flows before and after the occurrence of the unstable monkey saddle are very different. First, there are singular points north and south and afterward there are singular points east and west. The whole pattern seems to be rotated through a right angle. This transition is momentary and since the resulting patterns are so different the change is very dramatic. A similar transition is from the case shown in Figure 5(c) to that of Figure 5(d). The isolated trade areas in north and south are suddenly fused into one. Another example is provided by Figures 5(e) and (f).

The study of such sudden changes is catastrophe theory proper. As developed by Thom, this theory applies to gradient systems to a potential function. Fortunately, our flow lines represent gradient flows, so the theory is applicable. The object studied is the potential surface itself and the purpose is to classify all the catastrophes, or sudden changes of shape, that can occur with the potential surface when it undergoes a small deformation.

Without any restrictions at all, anything can happen. Therefore, a good procedure is again to use transversality. We are, however, not dealing with transversality and structural stability for potential surfaces any longer but for *families* of potential surfaces depending on a number of parameters. Transversality and structural stability used in this way yield the classification of elementary catastrophes when the number of variables and parameters is manageable.

If we dealt with the structural stability of potential functions, we would arrive at the conclusion that only Morse functions, i.e. functions with, at the most, some finite number of critical points with nonzero Hessians, would occur. The corresponding gradient flows would then be exactly those we have characterized as structurally stable, the hyperbolic singularities corresponding to the Morse critical points.

The use of transversality for families of potential surfaces would, on the contrary, admit such transitions as those illustrated in Figure 5. Moreover, if we suppose this new type of stability, every sudden dramatic change that could occur is completely classified up to topological equivalence. This is most remarkable.

In our case we deal with two variables, the two space coordinates, as was implicit in the discussion of potential surfaces. The catastrophes occurring then are not of the simplest types, the fold and cusp as encountered with one variable. We must start with umblic catastrophes, which are well classified for at least four parameter families. We would only need three parameters in the model. Only factors that influence the local transportation cost are relevant for the determination of the flow lines. In our specification of this transportation cost we used $f = r\kappa + w\lambda$. Now, the wage w depends on the flow lines and hence cannot represent an independent parameter. Capital rent r, on the contrary, can change over time because of capital accumulation. This process is exogenous to the model and so is the transportation technology, including the availability of roads. We must hence include the two technical coefficients κ, λ in the list of independent parameters. Together with capital rent they make three.

The corresponding catastrophes are accordingly of two possible types: the elliptic and hyperbolic umblic. The elliptic umblic has already been amply illustrated: it is nothing but the monkey saddle, or rather, the phenomena that can occur when it is disturbed. All these phenomena are confined to the class that can be obtained by varying the parameters in the expression

$$x^3 - 3xy^2 + a(x^2 + y^2) + bx + cy \tag{124}$$

which is called the universal unfolding for the function $x^3 - 3xy^2$. We have studied exactly these ways of disturbing the monkey saddle flow as an example. The remarkable fact is that this illustration covers half of all the phenomena of sudden change that can occur in the model.

The second half is represented by the hyperbolic umblic catastrophe. The function studied then is $x^3 + y^3$, with the universal unfolding

$$x^3 + y^3 + 3axy + bx + cy \tag{125}$$

This is easy to study by using simple geometry. The gradient flow has zero components in the coordinate directions when $ay = -x^2 - b/3$ or when $ax = -y^2 - c/3$. Geometrically, they are simple parabolas where a determines the sense and scale of the parabola and where b, c determine the intercepts. As the critical points are at the intersections of these two parabolas, directed at right angles, we can have no, two, or four critical points of the node and saddle types. The last case always comprises one source, one sink, and two saddles. Several of these structures may seem to be stable in the meaning used above. This, however, is deceptive. The patterns of hyperbolic singularities with no saddle connections are necessary in order that no catastrophes can occur, but only necessary and not sufficient.

20 CONCLUSIONS AND LOOSE ENDS

We can now summarize the arguments, point out some of the numerous loose ends, and hint at possible ways of completing the model. In Section 1 we conjectured that to a given locational structure for productive activity and for the distribution of residences there would correspond an optimum transportation pattern, and conversely. It was assumed that routes for transportation would be so chosen that their costs were minimized. They were so chosen, however, under the assumption that there was a given transfer cost function that represented the existing system of roads with respect to the geometric network shape and density distribution over the region. In a more general context this transfer cost function should be subject to optimizing choice. The author has worked with various ways of doing this (Puu 1979), but presently we just note that we have been little concerned by optimality.

It is even worse with respect to the converse conjecture, that to a given transportation system would correspond some location patterns with optimality or equilibrium characteristics. We have set down some relations belonging to a spatial general equilibrium model. By the specification of a production technology and the assumption of profit-maximizing behavior we have specified the forces behind the supply of goods and the

demand for labor services. However, about the other side of the market, which depends on the location of residences of workers, we have said very little. It is true that we have checked the general consistency of the general equilibrium system uncompleted as it stands, but otherwise we have only hinted that, in order that various locations should be equivalent for residential choice, the noncentral locations with resulting low wages and high prices must be endowed with compensating factors such as low housing costs due to low land rent. In this context, land rent was determined residually from productive activity and was proportional to the amount of capital invested. It is not quite satisfactory to disregard the use of land for housing and the effect of this on rent. In particular, the agglomeration of people with much capital invested for housing in a certain residential area should raise rent in the same way as the agglomeration of industrial activity does. We could also consider the effect of the demand for land for use in road building.

All this suggests that we should have introduced more types of production than just that of consumer goods, in particular the production of housing and transportation services. In addition, the decisions of the workers, landlords, and capitalists in choosing residential location should be introduced, or at least a condition of equivalence of various locations should be specified, in order that we arrive at a real general equilibrium model. As it stands, the present model just relates various flows of the Beckmann type by use of a production technology. This means that we actually have been little concerned with optima and equilibria, and have only hinted at location patterns that are just *compatible* with certain flows. The flows were assumed to be of the Beckmann type, i.e. they can be represented by vector fields that are related to the excess supply distributions by their divergences, and by flow lines obtainable from a gradient field to some potential. They were thus obtainable because it was assumed that the transportation system could be represented by a transfer cost field that was isotropic.

Despite the obvious abstraction from real networks, I am sure that these continuous models are valuable tools in regional modeling, too little explored as yet. Their advantage is that they make a number of powerful mathematical tools available to the economist. It is true that the more pictorial network models dealing with a finite set of nodes and edges arranged in graphs are increasingly yielding instruments that result from the development of mathematical programming and computation techniques. Such models are, however, not suitable for dealing with general geometric properties of cases that are not numerically specified. Thus the continuous models provide a complement.

The distinctive feature of the present study is the use of transversality to find which flows are structurally stable. It is surprising that this general principle of transversality, in conjunction with the weak assumptions that the flow is a gradient flow (a result of optimization) and that the potential is analytic (expandable in a power series), yields such rich results. First, we can conclude that because the analyticity makes critical points simple and because transversality makes them isolated the dynamic system in the neighborhood of a critical point behaves like a linear system of the simplest type. This reduces the possibilities for the types of critical points to those occurring with linear systems, i.e. nodes, saddles, and spirals. Second, the fact that the flows are potential flows rules out spirals. Since the most restrictive assumption we have made is that the directions of the flows are gradient directions to a potential, it is interesting that this assumption only rules out flows that wind around some point during infinite time without arriving at the goal during any finite period or that even circle around the critical point in closed orbits. Such flows of goods and labor services, or whatever, certainly make no economic sense.

We also assumed that unique prices and wages dominated at each location because of competition, so that the potential could be a single-valued function and the trajectories did not cross. Hence, we produce a system in which the flows do not stagnate along curve segments or on patches with areal content, but only at a number of isolated stagnation points that may be of two types: nodes and saddles. The location pattern around a node that is a sink or a source has circular ring symmetry (if the node is proper) and radial trajectories as in the classical von Thünen case. If the node is improper the trajectories become parabolic with a common tangent and the rings become elliptic. Only this latter case was seen to be compatible with structural stability, as the characteristic roots (eigenvalues) coincide by mere accident. Around a saddle the locational structure was such that the rings were split into sectors.

Transversality gave results even for the way in which the critical points could be fitted together. First, we saw that it is unlikely that two saddles are joined by a trajectory because there are only four trajectories, among the infinity of trajectories in the neighborhood of a critical point of saddle type, that go into it. Then, assuming the most general case that when going from one critical point along a characteristic line we sooner or later encounter a new critical point, we could arrange the set of critical points on a quadratic lattice. There is nothing odd in assuming that a trajectory from a saddle (one of the characteristic lines) ends at a node, since all trajectories in the neighborhood of a node are collected and incident to the critical point itself. It turned out that there was a unique way of orienting the graph consisting of the critical points and the quadratic grid joining them, where a node was always surrounded by four saddles and a saddle by four nodes. This resulted in a quadratic pattern of flows. By topological deformation a multitude of structurally stable flow and location patterns could be obtained. They were, however, all basically "quadratic" rather than "hexagonal" in shape.

To complete the picture, we tried to apply a flow orientation to a hexagonal graph; whenever this was done in a consistent way we obtained some monkey saddle, but monkey saddles were seen to be structurally unstable. The simplest and weakest perturbation would split them instead of introducing a smooth change in the location of the critical point and a small change of the exact shape of the trajectories around it, as is the case with perturbation of a structurally stable point. We concluded that in the hexagonal patterns there was an inherent structural instability. This was in strong contrast to the Christaller–Lösch tradition of hexagonal patterns, which are the closest possible approximations to the ideal circular trade areas. They are likely to turn up in considerations of optimum arrangements because of the extremum properties of the hexagonal tessellation that maximizes the market area of each firm, for a given length of the boundary enclosing it. In considerations of regular road networks too, a triangular tessellation (implying hexagonal subregions) emerges as the one that minimizes the detour factor.

It is not the intention here to suggest that the basic hexagonal shapes in theory should be replaced by quadratic ones. To assess the realism of the two patterns, empirical studies are needed. We can only point to the fact that hexagonal patterns are connected with flows that tend to have inherent instability.

We have already pointed to a number of loose ends left and building blocks lacking in a complete model. Even if this model is completed, certain questions remain. We have only considered one type of productive activity. What is the effect of introducing not only the production of housing and transportation services but also a great number of different productive activities possibly using several *different* transportation systems?

In addition, it would be most interesting to consider a model that includes a development over time when the system is not in equilibrium, involving migration flows and capital accumulation that may change the locational pattern of capital. A continuous dynamic system in three independent variables, time and the two space coordinates, would probably be the most yielding formulation as this may make the well investigated partial differential equations for diffusion and heat conduction or wave propagation applicable.

Another methodological question is whether it could simplify the analysis if we regard the region as a curved surface and the trajectories as geodesics. We broached the subject in the discussion of the potential surface in terms of differential geometry, but the question is whether a systematic use of tensors would simplify matters.

It cannot be overemphasized that the model is still incomplete as long as these questions are not answered and the lacking building blocks not supplied. Likewise, the use of the concept of structurally stable flows is tentative and rests on transversality considerations that tell us what is typical if all possibilities in some set are equally probable. However, when we know that something is true, transversality does not rule out facts even when they are improbable.

This study reflects two convictions of the author. The first is that aggregation from individual behavior relations derived from optimizing behavior are poor instruments when used alone in model building. At the micro level, optimizing behavior yields very few qualitative conclusions concerning the relations, and after aggregation the uncertainty is increased instead of reduced. In regional modeling the entropy model has introduced a most interesting aggregation method clearly superior to the methods envisaged by economists when aggregation problems were at issue in the 1950s. This is so because the use of the assumptions of independence of actions and *a priori* equiprobability of alternative choices reduces the uncertainty about macro relations as compared with micro. This is as things should be. Transversality is another similar modeling concept that, from a positive assumption of equiprobability when we have no knowledge about all factors, rules out certain cases as highly improbable.

The other conviction is that the continuous transportation models should have a power as scientific instruments far beyond the results they have yielded in pure transportation analysis. The flow property and not only the continuous transfer cost field property may be useful. Even though computation algorithms and computation efficiency and cost have developed amazingly, classical mathematical methods that have been perfected over centuries in applications to physical problems still supply a reservoir of unexploited methods.

APPENDIX GENERICITY AND STRUCTURAL STABILITY

I have tried to make the exposition in terms of classical calculus and the classical theory of differential equations, with slight reference to transversality. However, the treatment would be very incomplete if nothing were said about the global theory of differential equations, which yields results that can be used in a more direct way. The problem with this approach is that by employing methods of differential topology and concentrating on manifolds with a much more complicated and general nature than the plane, it uses a terminology that is too general for the present purpose. At the same time it is difficult to

gain an intuitive understanding of the facts involved when the highly general and abstract methods of topology are used.

To complete the picture, however, a brief survey will be given of the results needed from the theory of differentiable flows on manifolds. The line of development was opened by Poincaré and Birkhoff. The results needed were announced in 1937 by Pontryagin and Andronov, but the complete development is due to Smale (1967) and Peixoto (1973).

By a *generic property* of an operator or of a vector field is meant that among the set of all operators or vector fields the subset characterized by a generic property is dense and open. That is, any generic or nongeneric vector field at a point can be approximated arbitrarily closely by a generic field, whereas the converse is not true. From this it can be suspected that nongeneric properties are associated with instability, as the generic class is reached as soon as the system is changed a little, whereas generic properties are associated with stability. In fact it is demonstrated that *structural stability* is a generic property of vector fields. Structural stability means stability to perturbation, i.e. small changes of the structure of the dynamic system.

A *perturbation* is usually explained in the following way. Let us consider two vector fields $f(x)$ and $g(x)$ with f,g: $R^2 \to R^2$. We define a norm (Euclidean or other) on the components of the difference between the mapped vectors, $f(x) - g(x)$, and denote it by $|f(x) - g(x)|$. Moreover, we define a norm on the difference between the operators, $Df(x) - Dg(x)$, as $\max \{|(Df(x) - Dg(x))x| : |x| \leqslant 1\} = ||Df(x) - Dg(x)||$ where D is the derivative conceived as a 2×2 matrix. If the vector x takes values on the unit disk then the two mappings transfer the vector to the positions $Df(x)x$ and $Dg(x)x$, respectively. We apply the previously used norm to measure the distance between the two images. The norm on the operator is hence the maximum distance between the images into which a point on the unit disk is mapped. If we now suppose that $|f(x) - g(x)| < \epsilon$ and $||Df(x) - Dg(x)|| < \epsilon$, we make sure that the differential equations $\dot{x} = f(x)$ and $\dot{y} = g(x)$ are "close" or, in other words, perturbations of each other. If a perturbation does not alter the character of the trajectories, so that we can find a homeomorphism (i.e. a continuous one-to-one mapping) $R^2 \to R^2$ that maps trajectories for the g system on to trajectories of the f system, then the latter is said to be structurally stable. This is a precise definition of structural stability.

We are now ready to turn to what is typical for structurally stable systems. By a *hyperbolic singularity* or *critical point* is meant a critical point where the real parts of all eigenvalues are nonzero. This concept should not be confused with the concept of a hyperbolic point on a surface such as a potential surface. There, a hyperbolic point meant a saddle, whereas a maximum or a minimum was termed an elliptic point. In the present context a hyperbolic point can be a sink, a source, or a saddle. For gradient flows the singularities are nodes, since spirals do not occur. The *index* of a hyperbolic singularity designates how many eigenvalues are positive and how many are negative.

An important result is that a hyperbolic critical point is transferred close to its original position and retains its index by a perturbation, provided that Df is an invertible mapping. However, this condition only means that the Jacobian of the system of differential equations (or the Hessian of the potential) is nonzero. With this provision, nodes and saddles are stable to perturbation.

Another important result is that no trajectory joins two saddles.

REFERENCES

Angel, S., and G.M. Hyman (1970) Urban velocity fields. Environment and Planning 2:211–224.
Angel, S., and G.M. Hyman (1972) Urban spatial interaction. Environment and Planning 4:99–118.
Angel, S., and G.M. Hyman (1976) Urban Fields – A Geometry of Movement for Regional Science (London: Pion).
Beckmann, M. (1952) A continuous model of transportation. Econometrica 20:643–660.
Beckmann, M. (1953) The partial equilibrium of a continuous space market. Weltwirtschaftliches Archiv 71:73–89.
Beckmann, M. (1968) Location Theory (New York, NY: Random House).
do Carmo, M.P. (1976) Differential Geometry of Curves and Surfaces (Englewood Cliffs, NJ: Prentice–Hall).
Cartan, H. (1963) Elementary Theory of Analytic Functions of One or Several Complex Variables (Reading, MA: Addison–Wesley).
Fejes Tóth, L. (1964) Regular Figures (Oxford: Pergamon).
Fox, C. (1954) Introduction to the Calculus of Variations (London: Oxford University Press).
Frisch, R. (1965) Theory of Production (Amsterdam: North-Holland).
Isard, W. (1956) Location and Space Economy (Cambridge, MA: MIT Press).
Kantorovich, L. (1958) On the translocation of masses. Management Science 5:1–4.
Marsden, J.E. (1973) Basic Complex Analysis (San Francisco, CA: Freeman).
Marsden, J.E., and A.J. Tromba (1976) Vector Calculus (San Francisco, CA: Freeman).
Morse, M. (1934) The calculus of variations in the large. American Mathematical Society Colloquia Publications 18.
Peixoto, M.M. (ed.) (1973) On the classification of flows on 2-manifolds. In: Dynamical Systems (New York, NY: Academic Press).
Poston, T., and I. Stewart (1978) Catastrophe Theory and its Applications (London: Pitman).
Puu, T. (1977) A proposed definition of traffic flow in continuous transportation models. Environment and Planning 9:559–567.
Puu, T. (1978) On the existence of optimal paths and cost surfaces in isotropic continuous transportation models. Environment and Planning 10:1121–1130.
Puu, T. (1979) The Allocation of Road Capital in Two-Dimensional Space (Amsterdam: North-Holland).
Simmons, G.F. (1972) Differential Equations (New York, NY: McGraw-Hill).
Smale, S. (1967) Differentiable dynamical systems. Bulletin of the American Mathematical Society 73:747–817.
Thom, R. (1969) Topological models in biology. Topology 8:313–335.
Wardrop, J.G. (1969) Minimum cost paths in urban areas. Strassenbau- und Strassenverkehrstechnik 86:184–190.
Warndtz, W. (1967) Global science and the tyranny of space. Papers and Proceedings of the Regional Science Association 19:7–19.

LIST OF SYMBOLS

- A land as productive input
- C aggregate consumption value
- c cost of transporting a quantum of goods or labor from one location to another
- f cost of transfer across a point
- G aggregate land rent value
- g local land rent
- K capital as productive input
- k capital density per unit land area
- L labor as productive input

l, l'	labor demand, supply per unit land area
l_1, l_2	components of labor flow field
$\boldsymbol{l}$	flow of labor services
$\|\boldsymbol{l}\|$	flow intensity for labor services
$\nabla . \boldsymbol{l}$	excess supply of labor
M	aggregate import value
M_l	aggregate labor import value
M_q	aggregate goods import value
$\boldsymbol{n}$	unit normal to the boundary of the region
p	local price of goods
Q	productive output of goods
q, q'	goods supply, demand per unit land area
q_1, q_2	components of goods flow field
$\boldsymbol{q}$	flow of goods
$\|\boldsymbol{q}\|$	flow intensity for goods
$\nabla_{\circ} \boldsymbol{q}$	excess supply of goods
R	aggregate capital rent value
$\mathcal{R}$	the region considered
$\partial\mathcal{R}$	boundary of the region
r	local capital rent
s	arc length parameter
T	aggregate transportation cost
T_l	aggregate labor transportation cost
T_q	aggregate goods transportation cost
U	individual utility of per capita consumption
$(u,v), (x,y)$	Cartesian space coordinates
W	aggregate wage incomes
w	local wage rate
X	aggregate export value
X_l	aggregate labor export value
X_q	aggregate goods export value
α, β, γ	coefficients in Cobb–Douglas production function
κ_1, κ_2	roots of characteristic equation
λ	Lagrangian potential function
$\nabla\lambda$	gradient field
μ	coefficient in Minkowski metric
(ρ, ω)	polar space coordinates
σ	"natural" arc length parameter

THE AUTHOR

Tönu Puu has been Ordinary Professor and Chairman of the Department of Economics at Umeå University in Sweden since 1971. His present work is on continuous spatioeconomic modeling.

Professor Puu graduated in 1959 from Uppsala University and received his Ph.D. in economics from there in 1964. His dissertation dealt with mathematical models for portfolio selection. His later research publications deal with investment theory, production theory, the methodology of social science, the management of natural resources, transportation, and general spatial economics.

RELATED IIASA PUBLICATIONS

Research Reports

Andersson, A.E., and H. Persson (1980) Integration of transportation and location analysis: A general equilibrium approach. RR-80-40 (available for a handling charge of $1.00).
Reprinted from *Papers of the Regional Science Association* vol. 42 (1979).

Leonardi, G. (1981) A unifying framework for public facility location problems. RR-81-28 (available for a handling charge of $1.00).
Reprinted from *Environment and Planning* A, vol. 13 (1981).

Leonardi, G. (ed.) (1982) Public facility location: Issues and approaches. RR-82-23 (available for a handling charge of $3.00).
Reprinted from *Sistemi Urbani* vol. 3 (1981).

Snickars, F., and A. Granholm (1982) A multiregional planning and forecasting model with special regard to the public sector. RR-82-21 (available for a handling charge of $1.00).
Reprinted from *Regional Science and Urban Economics* vol. 11 (1981).

Tobler, W. (1975) Spatial interaction patterns. RR-75-19 ($5.00).