SYSTEM DYNAMICS REPRESENTATION OF CATASTROPHE AND ITS APPLICATION TO TRANSPORTATION

by

Jiefeng Qin

Thesis submitted to the Faculty of the

Virginia Polytechnic Institute and State University

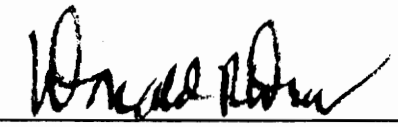
in partial fulfillment of the requirements of the degree of

Master of Science

in

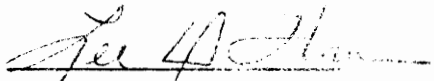
Civil Engineering

APPROVED:



Dr. Donald R. Drew, Chairman

Dr. Antonio A. Trani



Dr. Lee D. Han

April 27, 1992

Blacksburg, Virginia

System Dynamics Representation of Catastrophe And Its Application To Transportation

by

Jiefeng Qin

Committee Chairman: Dr. Donald R. Drew

(Abstract)

For a long time mathematicians have been developing a number of theorems that seek to establish general structural and behavioral characteristics of dynamic systems. Most of these techniques based on the calculus have been designed for the study of continuous phenomena. Hence they are ineffective to deals with discontinuous or divergent behaviors. Catastrophe theory, when applied to scientific problems, deal with the properties of discontinuities directly without reference to any specific underlying mechanism. It is especially suited to the study of systems in which the only reliable observations are of the discontinuities.

System dynamics, introduced by professor Jay W. Forrester in the early 1960's, is used to represent general, complex dynamic systems. It focuses on the structure and behavior of systems composed of interacting feedback loops. The

nature of its approach to modeling shares many common points with catastrophe theory. Particularly, both are used to seek to develop fruitful simplifications of a complex reality.

The purposes of this thesis, therefore, are: first, to offer a qualitative as well as a quantitative description of catastrophe theory, as this theory is not very familiar to many people; secondly, to present the relationship between catastrophe theory and system dynamics; and thirdly, to apply these theorem to urban transportation planning.

Abstract

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to those individuals who provided assistance in the preparation of this thesis. I am most grateful to my advisor Dr. Donald R. Drew, for his patience and persistence in directing this research. Appreciation is also expressed to the other members of the committee: Dr. Antonio A. Trani, and Dr. Lee D. Han.

CONTENTS

1. CATASTROPHE THEORY

1.1 A General Description

A great many of the most interesting phenomena in nature involve discontinuities. These may be in time, like the abrupt bursting of a bubble, or they may be spatial, like the change in an object's shape. Yet most mathematical principles available are ideally suited to analyze smooth, continuous, quantitative changes. These techniques, based primarily on the calculus, represent a state of some systems of interest as an equilibrium point which is a function of some parameters or independent variables. When the parameters change slowly and smoothly, then the position of the equilibrium point also changes slowly and smoothly. They have so far been proved ineffective to deal with the discontinuous, divergent and conflicting phenomena. This is illustrated by using an example about the behavior of the dog in Zeeman's book "Catastrophe Theory" [15]. If only one of two emotional factors, rage or fear, is present, the response of the dog is easy to predict. If the dog is frightened then it will most likely retreat. But rage will lead it to attack. What if there are both rage and fear emotions simultaneously existing in the dog's feeling? Classic models which cannot accommodate the discontinuity might predict a neutral behavior because of the compensation of

these two conflicting stimuli. It simply reveals the shortcomings of such models since the neutral behavior is the least probable behavior. When the dog is made to feel both angry and fear, the probabilities of both extreme modes are high. The dog will either attack or flee instead of remaining indifferent in most cases. It is, therefore, necessary to develop a new idea to account for this phenomenon.

Catastrophe theory is a controversial new way of thinking about change -- change in a course of events, change in a system's behavior, change in ideas themselves. Its name suggests disaster, and indeed the theory can be applied to literal catastrophes such as the collapse of a bridge or the slump of the stock market. But it also deals with changes as quiet as the moving of sunlight and as subtle as the transition from ice to water.

As a part of mathematics, catastrophe theory grows where algebra, calculus, and topology meet. It is a mathematical method for dealing with discontinuous and divergent phenomena directly, without reference to any specific underlying mechanism. The theory has the potential for describing the solution of forms in all aspects of nature, and it is concerned with sudden and discrete changes in system state variables resulting from a slow, smooth and small change in one or more parameters.

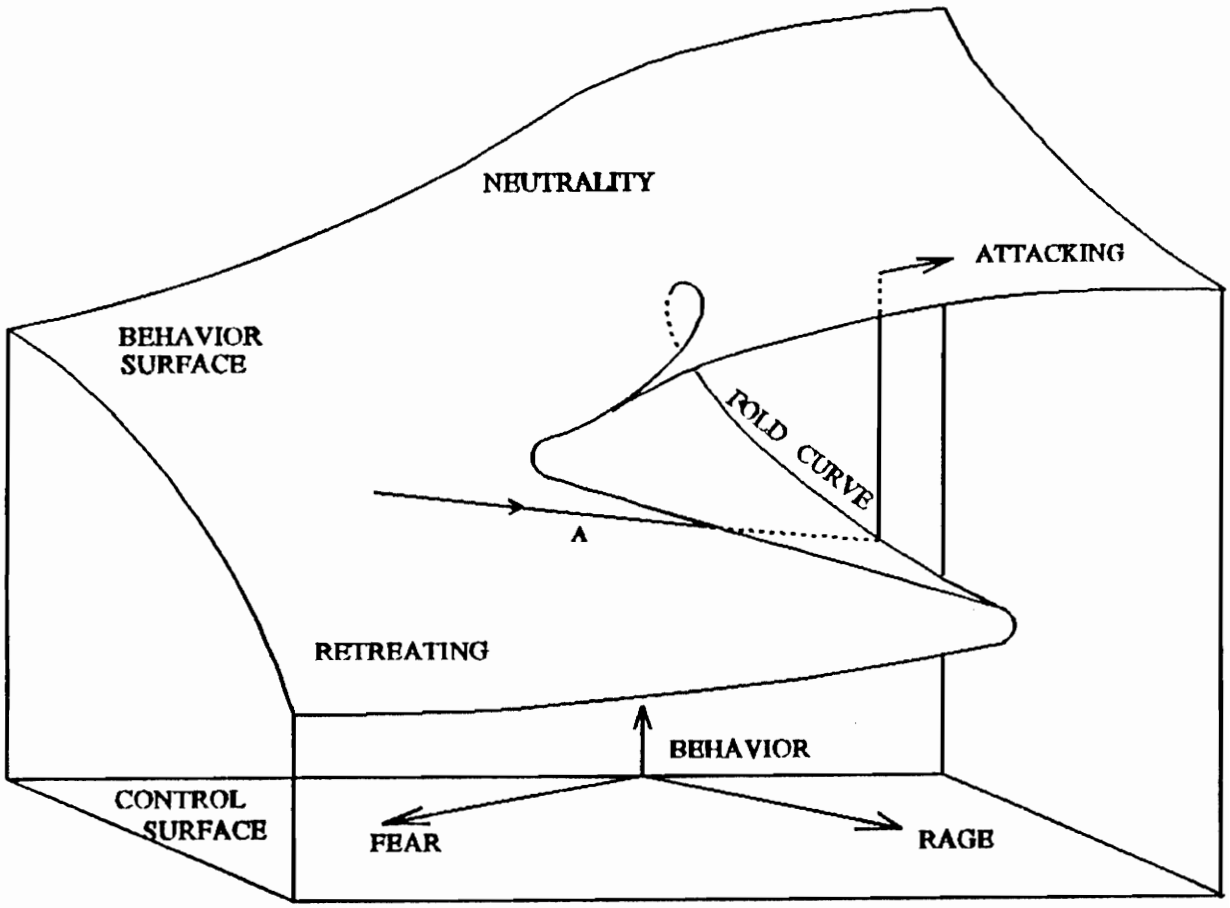


Figure 1 Cusp catastrophe model of a dog's behavior

Based on one of the elementary catastrophes, the possible behavior of the dog in the previous example can be predicted under any particular circumstances. To construct the model Zeeman displayed a special diagram in three dimensional space. This is plotted in Figure 1. There are two axes representing fear and rage. The behavior of the dog, which ranges from attacking to retreating, is measured on a vertical axis. For any combinations of rage and fear, there is at least one

likely form of behavior, measured by appropriate height on the vertical axis. But at the middle of the diagram where rage is roughly equal fear, there are two sheets representing either attacking or fleeing. These are connected by a pleat which represents the neutrality, a least likely behavior. The line defined the edge of the pleat is called fold curve. If an frightened dog is made more angry, its mood follows the trajectory A on the control surface. The corresponding behavior moves to the right of the bottom sheet until it reaches the fold curve; then the path must jump suddenly to top sheet as the bottom one vanishes. This abrupt change in behavior might be called catastrophe. It is the strength of the model derived from catastrophe theory that it can account for this sudden, discontinuous change while the independent parameters move smoothly and slowly.

Catastrophe theory is the invention of Rene Thom. He presented his idea in a book named Structural Stability and Morphogenesis which was published in 1972. The theory is derived from topology, the branch of mathematics concerned with the properties of surfaces in many dimensions. For catastrophe theory, the problem is to describe the shape of all possible equilibrium surfaces. It can be solved in terms of a few archetypal forms, which are called the elementary catastrophes. Thom has shown that there are just seven elementary catastrophes for processes controlled by no more than four factors.

1.2 The Nature of Catastrophe Theory

The underlying mathematics of catastrophe theory is based on topologic methods, and this takes it beyond the control of those with a relatively elementary knowledge in classical calculus. This is certainly the essence on which the theory relies.

Since the bifurcation theory, a study of the "bifurcation" of new solutions to an equation or system of equations from some known solutions as a parameter changes, is relevant to systems of a more general type than the catastrophe theory, it is useful to develop the argument in this direction. Indeed, this kind of theory is the most important in most applied work. Most of the illustrations fall within the program of catastrophe theory. This limits systems of interest to so-called gradient systems which arise from the minimization of some objective functions and associated dynamics.

If x_i is one of a set of state variables describing a system, or one dependent variable to be predicted in a model, and independent variable u_i is control variable, then the first order equations of motion in a gradient system can be derived from a potential function V by

$$\frac{dx_i}{dt} = -\frac{\partial V}{\partial x_i} \qquad (1)$$

The equilibria of the system is determined by the equations

$$\frac{dx_i}{dt} = 0 \quad \text{or} \quad \nabla V = \frac{\partial V}{\partial x_i} = 0 \qquad (2)$$

where the minimum of the potential function V occurs. The appearance of the gradient of the potential function V explains the name of this type of system. Many systems can in fact be defined in this way even though it is not obvious at first sight. The study of the equilibrium positions of a gradient system, and how these equilibria move about and change characters while the control parameters change, is called elementary catastrophe theory.

For any given independent variables u, there are a number of corresponding equilibrium points given by the solutions of equation (2) which are the minima of the potential function in equation (1). These points determine a surface in the space (x,u) and represent possible equilibrium states of the system.

The classic mathematics of equations (1) and (2) is well known. As one or some of u variables change slowly and smoothly, a corresponding smooth move

in the state variables x can be predicted. It can be seen explicitly that for this to occur, the surface defined by the equilibrium solutions has to be itself smooth and not folded in any way. For any combinations of u variables, if there are multiple solutions for x, then something more complicated will takes place. What Thom, and others, have done is to deal with these complications, and to prove that they fall into a small group of basic types.

The solutions to equation (2) describe the stationary points of the potential function V or, more precisely, of a family of functions of x, expressed in the form of u. Normally, stationary points are maxima or minima which are distinguished by the Hessian matrix of second derivatives being negative or positive respectively, or by the second derivative of V if there is only one state variable. While the Hessian matrix of second derivatives is singular or the second derivative of V equals zero, the stationary points are not maxima or minima. They are called singularities and it is well-known that unusual system behavior can be observed and anticipated at and near these points. The catastrophe theory classifies these kinds of singularities which can occur. If, for example, there is a single state variable and two control variables, x and (u_1, u_2), respectively, the surface of steady-state points around a singularity will be equivalent, in a topological sense, to a cusp surface in three dimensional space. This is plotted in Figure 2.

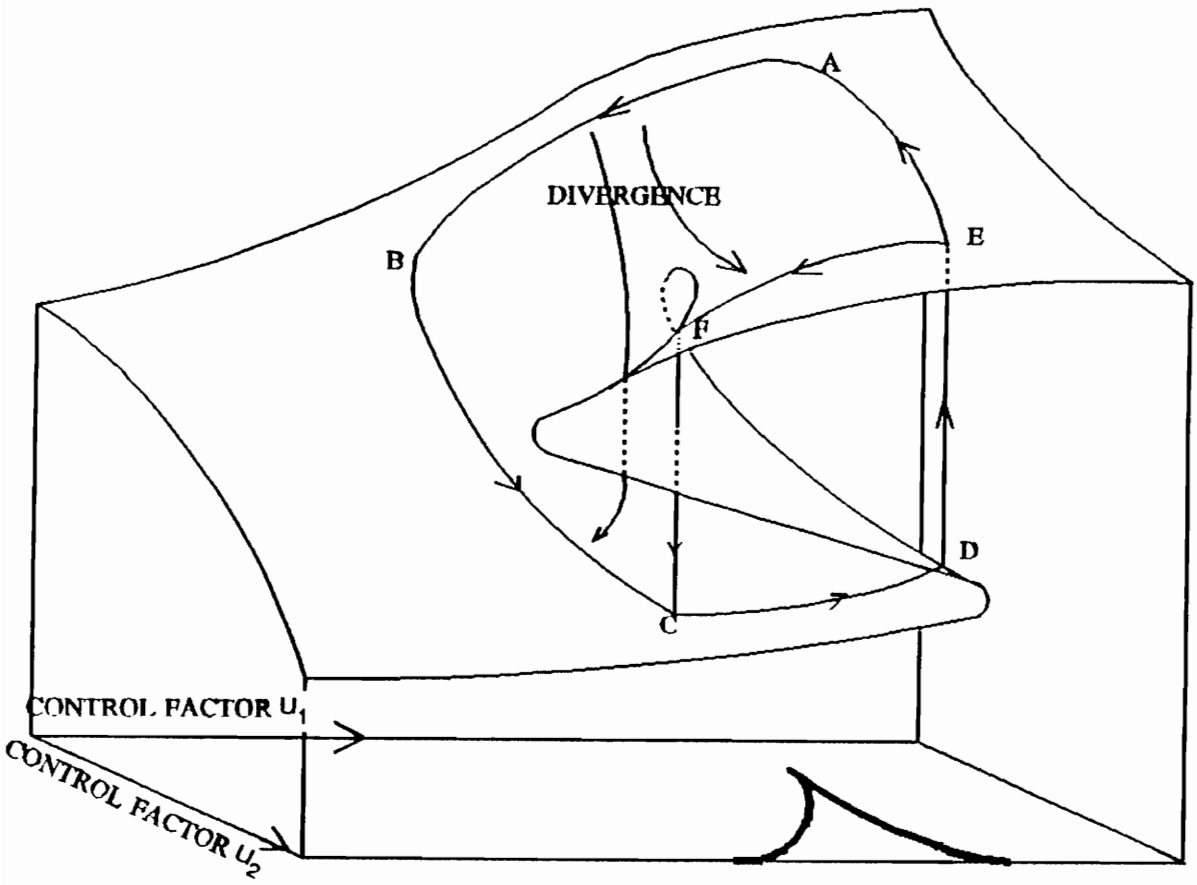


Figure 2 Illustration of catastrophe, bifurcation and divergence

In effect, "in a topological sense" means that the surface defined by possible equilibrium values for a system can be forced into the form of Figure 2 after some transformations of variables where necessary. The achievements of the appropriate transformations in real work are often likely to be a very difficult task, although insights can be gained without it being carried through explicitly.

Now it is important to discuss the heart of the matter: catastrophe. The possibilities of catastrophe are illustrated in Figure 2. One of the most distinctive features is that there is a smooth double fold in the middle of the surface, creating a non-crease pleat, which grows narrower from the front of the surface to the back and eventually disappears in a singular point where the three sheets of the pleat come together. As noted earlier, every point, with the exception of those on the middle sheet, on the surface represents a possible behavior of the equilibrium state. All the points along the fold curve, which form the "lip" on the pleat, are semi-stable points of inflection. All the rest points are stable minima. The middle sheet represents the least possible actions or unstable maxima.

For some combinations of values of the control parameters, u_1 and u_2, there appear to be three corresponding state solutions, one on the upper surface of the pleat, one on the bottom surface and one on the pleat. It turns out that the points on the top and bottom sheets are stable points since the potential function V at these points is minimum, while the third possibility, the points on the pleat, is unstable or inaccessible. This means if the system occupies this state, any disturbance will force it to the stable point above or below. So only two stable states exist under some conditions, though it looks like there is a third possibility. It is intuitive that the function V is singular at the points on the fold curve where the stable changes to unstable, or vice versa. Now, the change of the system

behavior can be visualized by a point moving along the surface composed of equilibrium points, or so-called behavior surface. Select two points near the axis of control factor u_1 which are slightly different in control variable u_1 but at the same value of u_2. As the value of control factor u_2 then increases, the points trace parallel paths to the front of the surface. If one travels onto the top surface while the other onto the bottom, instead of both of them passing on the same side of the pleat, then the divergent behavior of the system occurs. This depends on the precise value of control variable u_1 as the moving point passes the beginning of the pleat.

The divergent paths are still smooth changes in behavior. But the catastrophe graph indicates the probabilities of discontinuous changes, those which occur when a point moving to the left or right reaches the lip of the pleat. The system can move smoothly from A to B, B to C, and vice versa. But if the control variable u_1 is increased with the system at point C, the behavior point reaches D and there is nowhere else to go. At this time, the stable point has to turn into a semi-stable point of inflection, and any further increase in control factor u_1 obliges the system to jump to the stable minimum E. The system passes the non-equilibrium states as quickly as possible; the transition is catastrophic. A same jump can take place if the behavior point at E is changed by the decrease of factor u_1: it moves smoothly to F, then suddenly and catastrophically jumps to

point C.

If the folded region is projected vertically downwards onto the u-plane, a cusp-shaped section of that plane is obtained. This contains a set of values of u known as bifurcation set, which are in some sense critical: outside the cusp region, the system only has one steady-state available to it; inside there are two possible stable states; as the boundary of the critical region is crossed, jumps can take place. This kind of catastrophe is called cusp catastrophe, one of the simplest forms of catastrophes.

1.3 Several Features of Cusp Catastrophe

The cusp catastrophe shows a number of qualitative features which are interrelated. They are:

1. Sudden jump or catastrophe

This is the most distinguished phenomenon in cusp catastrophe graph. As the behavior trajectory reaches the fold curve, an abrupt jump is unavoidable.

2. Bimodal

Under some circumstances there are two equilibrium states and hence possible conflict.

3. Divergence

A small difference in approach leads the system to very different states as the trajectory shown in Figure 2.

4. Hysteresis

The cusp catastrophe graph in Figure 2 shows that it is possible to get from C to E either by a smooth path or via a catastrophe. Which will happen in any special cases depends on the extent and the sequence of the changes in the control variables. If, for example, a system is at C and control factor u_1 alternately increases and decreases by a suitable amount, the behavior of the system seems to be a cycle with two smooth portions linked by catastrophes.

5. Delay convention

Inside the bifurcation set, the behavior of the system is determined by a delay convention, a rule supplied to determine which of the multiple possibilities the system adopts. The two most common are firstly, perfect delay, which means that the system stays in its original state until that state disappears as the

trajectory leaves the bifurcation set; and secondly, the Maxwell convention, which assumes that if more than one minimum is available, the system chooses the state which represents the lowest.

1.4 Mathematics of Cusp Catastrophe

In order to provide a concrete example in which all the variables are obvious and measurable, and the relationship between them differentiable and computable, a short description is offered of a catastrophe machine. This small educational toy invented by E.C. Zeeman is made out of elastic bands designed to illustrate catastrophe theory, and especially the cusp catastrophe. It consists of a disc mounted flat against a board, able to turn freely. To one point Q on its edge is attached two lengths of elastic bands. One of these bands has its other end fixed to the board at point R, far enough from the hub O of the wheel to keep the elastic bands QR always tight. The second has its other end P free to move.

The distance to be measured is taken to make the diameter of the disc equal to 1, the lengths of unstretched rubber bands equal to 1, and the distance OR equal to 2.

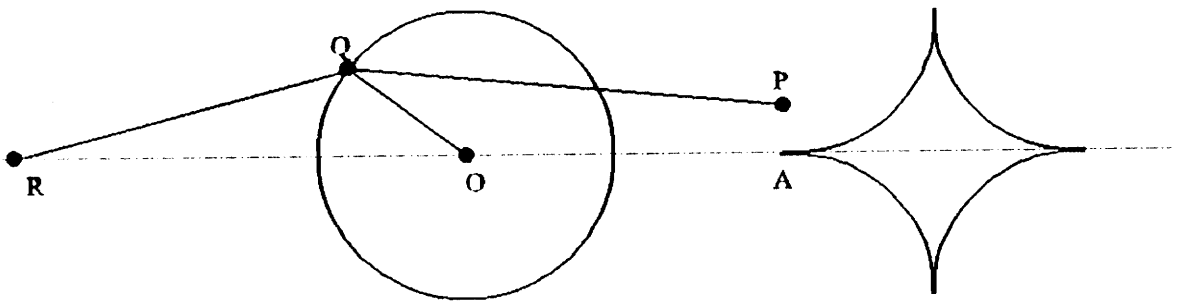


Figure 3 Catastrophe Machine [13]

Now the catastrophe machine can be operated by moving the control point P slowly and smoothly in the plane of the disc. In many cases the movements of control point P cause only smooth rotations of the disc, but in some cases the disc swings suddenly from one side to the other. If all the positions of P at which these jumps occur are marked on base, a diamond-shaped curve is generated. This curve is made up of four connected cusp-shaped curves, the bifurcation sets of four cusp catastrophes.

If only one of the four cusps is considered, the corresponding behavior surface can be constructed by arranging the fold curve directly over the cusp. For any position of the control point P outside the cusp the behaviors of the system have only one corresponding stable state. If the control point P falls into the cusp, there are three sheets of behavior surface corresponding to different system behaviors, but again the middle one is to be excluded.

In principle, the analysis of the "catastrophe machine" is perfectly straightforward. What is happening is that the position of the disc is determined by the minimum potential energy of the two elastic bands. The state of the system at any time can be specified by a single variable θ, the angle that the line OQ makes with the symmetric axis. Since the energy of a stretched elastic is proportional to the square of its extension, the potential energy of the system is

$$V(\theta) = \frac{1}{2}k[(l_1-1)^2 + (l_2-1)^2] \tag{3}$$

where l_1 and l_2 are the lengths of the rubber bands after they have been stretched and k is their modules of elasticity.

Considering a special case in which P is only allowed to move along the symmetric axis, then, by assuming the distance OP to be r,

$$l_1^2 = r^2 + \frac{1}{4} + r\cos\theta \tag{4}$$

$$l_2^2 = 4 + \frac{1}{4} - 2\cos\theta \tag{5}$$

and

$$l_1 = \sqrt{r + \frac{1}{4} + r\cos\theta} \tag{6}$$

$$l_2 = \sqrt{4 + \frac{1}{4} - 2\cos\theta} \qquad (7)$$

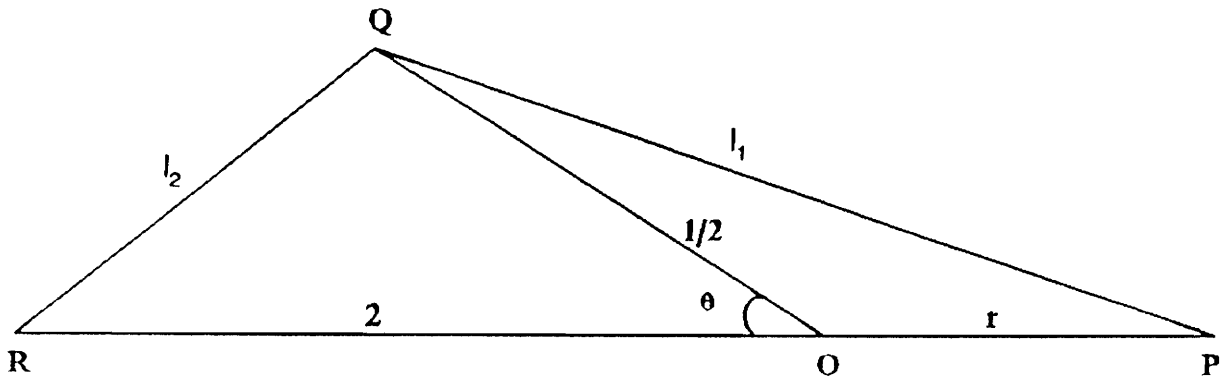


Figure 4 Dimensions of catastrophe machine for control point on symmetric axis

It is clear that as point P runs along the symmetric axis, there will always be an equilibrium position at $\theta = 0$, such as A in Figure 3, by general principles of static. Whether it is stable or unstable can be explored by taking the second derivative of the function V. Since it is difficult to obtain results from the above equations which turn out to be awkward to work with, it is necessary to expand potential function V as a series of θ.

By using Taylor series at $\theta_0 = 0$ and up to terms of order four, equation (4) and (5) become

$$l_1^2 = r^2 + \frac{1}{4} + r\left(1 - \frac{\theta^2}{2} + \frac{\theta^4}{24}\right) + O(5) \tag{8}$$

and

$$l_2^2 = \frac{17}{4} - 2\left(1 - \frac{\theta^2}{2} + \frac{\theta^4}{24}\right) + O(5) \tag{9}$$

Since at $\theta_0 = 0$

$$l_1 = r + \frac{1}{2}$$

$$l_1^{(1)} = 0$$

$$l_1^{(2)} = -\frac{r}{2}\left(r + \frac{1}{2}\right)^{-1}$$

$$l_1^{(3)} = 0$$

$$l_1^{(4)} = \frac{r}{2}\left(r + \frac{1}{2}\right)^{-1} - \frac{3r^2}{4}\left(r + \frac{1}{2}\right)^{-3}$$

equation (6) can be expressed as:

$$l_1(\theta) = l_1(0) + \theta l_1^{(1)}(0) + \frac{\theta^2}{2!} l_1^{(2)}(0) + \frac{\theta^3}{3!} l_1^{(3)}(0)$$

$$+ \frac{\theta^4}{4!} l_1^{(4)}(0) + O(5)$$

$$= (r+\frac{1}{2})\,[1 - \frac{r\theta^2}{4(r+\frac{1}{2})^2} + \frac{\theta^4}{16}(\frac{r}{3(r+\frac{1}{2})^2} - \frac{r^2}{2(r+\frac{1}{2})^4})] \quad (10)$$

And similarly equation (7) becomes

$$l_2(\theta) = \frac{3}{2} + \frac{1}{3}\theta^2 - \frac{7}{108}\theta^4 + O(5) \quad (11)$$

Substituting equation (8), (9), (10) and (11) into (3) yields

$$V(\theta) = \frac{1}{2}k[(r-\frac{1}{2})^2 + \frac{1}{4} + (\frac{1}{3} - \frac{r(2r-1)}{2(2r+1)})\theta^2$$

$$+ (\frac{r}{24} - \frac{r}{12(2r+1)} + \frac{r^2}{2(2r+1)^3} + \frac{5}{108})\theta^4] + O(5) \quad (12)$$

The first two derivatives of V to θ are accordingly

$$\frac{dV}{d\theta} = k[(\frac{1}{3} - \frac{r(2r-1)}{2(2r+1)})\theta + 2(\frac{r}{24} - \frac{r}{12(2r+1)}$$

$$+ \frac{r^2}{2(2r+1)^3} + \frac{5}{108})\theta^3] + O(5)$$

$$\frac{d^2V}{d\theta^2} = k[(\frac{1}{3} - \frac{r(2r-1)}{2(2r+1)}) + 6(\frac{r}{24} - \frac{r}{12(2r+1)}$$

$$+ \frac{r^2}{2(2r+1)^3} + \frac{5}{108})\theta^2] + O(5)$$

The nature of the equilibrium is determined by the sign of the second derivative $d^2V/d\theta^2$ at $\theta = 0$, i.e.

positive if $\frac{1}{3} > \frac{r(2r-1)}{2(2r+1)}$

negative if $\frac{1}{3} < \frac{r(2r-1)}{2(2r+1)}$

It follows that the equilibrium changes from a maximum to a minimum or vice versa where

$$\frac{1}{3} = \frac{r(2r-1)}{2(2r+1)}$$

that is

$$6r^2-7r-2 = 0$$

The solutions to this equation are

$$r = \frac{7 \pm \sqrt{97}}{12}$$

The negative root is extraneous, because it corresponds to a position of P for which the rubber bands are not both stretched. However, the positive root

$$r = \frac{7 + \sqrt{97}}{12} \approx 1.40$$

specifies the position of A.

At point A, the real root of $dV/d\theta$ is $\theta = 0$, but $d^2V/d\theta^2$ and $d^3V/d\theta^3$ both vanish. As the sign of the fourth derivative is positive, the equilibrium at A is stable.

In general when θ is small, the potential energy function V can be expressed as

$$V(\theta) = a + b\theta^2 + c\theta^4$$

Taking the first two derivatives yields

$$\frac{dV}{dt} = 2b\theta + 4c\theta^3$$

$$\frac{d^2V}{dt^2} = 2b + 12c\theta^2$$

If P is at the left side of A, which means P is outside the cusp, then both b and c are positive. The real root of the equation $dV/d\theta$ is $\theta = 0$. And since the $d^2V/d\theta^2 > 0$, the equilibrium is stable. When P is moved into the cusp, the coefficient b becomes negative. As a result, the equation $dV/d\theta = 0$ has three real roots, $\theta = 0$ and $\theta = \pm\sqrt{(b/2c)}$. As the second derivative is checked, the equilibrium corresponding to $\theta = 0$ is unstable and those corresponding to the others are stable.

Since A is located, the behavior of system when P is near A can be analyzed in detail. Considering the following case in Figure 5 where P is going around A, the formula for l_2 is as before; whereas

$$l_1^2 = (r + \alpha + \frac{1}{2}\cos\theta)^2 + (\frac{1}{2}\sin\theta - \beta)^2$$

This gives, to fourth order in θ,

$$l_1^2 = (r+\alpha)^2 - \beta\theta - \frac{1}{2}(r+\alpha)\theta^2 + \frac{1}{6}\beta\theta^3 - \frac{1}{24}(r+\alpha+1)\theta^4 + \frac{1}{16}\theta^4$$

and

$$l_1 = (r+\alpha+\frac{1}{2}) - (r+\alpha+\frac{1}{2})^{-1}[\frac{1}{2}\beta\theta+\frac{1}{4}(r+\alpha)\theta^2-\frac{1}{12}\beta\theta^3-\frac{1}{48}(r+\alpha)\theta^4]$$

$$-\frac{1}{8}(r+\alpha+\frac{1}{2})^{-3}[\frac{1}{4}(r+\alpha)^2\theta^4+r\beta\theta^3]$$

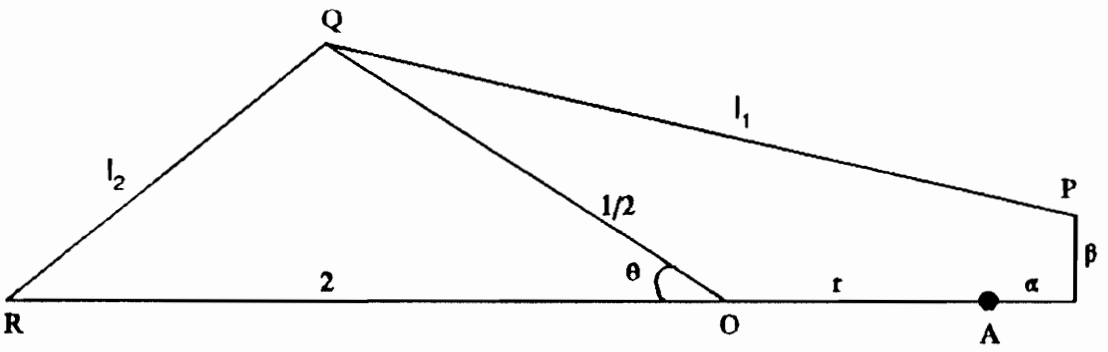


Figure 5 Dimensions of catastrophe machine

It is worth noting that in both I_1 and I_2, α is a factor of the coefficients of the even powers of θ but not of the odd ones, whereas β occurs only in the coefficients of odd powers. It is known from the previous calculation that when $\alpha = \beta = 0$ the coefficients of θ^2 vanish but those of θ^4 do not. So if the corresponding energy V is worked out, the result can be expressed of the form

$$V(\theta) = \frac{1}{2}k(a_0 + a_1\beta\theta + a_2\alpha\theta^2 + a_3\beta\theta^3 + a_4\theta^4 + O(5)) \qquad (13)$$

where $a_0, \cdots, a_4$ are certain constants. These values are not needed for the present discussion although they can be evaluated by explicit calculations.

Note at this stage that at A, where $\alpha = \beta = 0$, the V is of the form $a_4\theta^4 +$ O(5), and hence A is a degenerate critical point.

A much deeper theorem shows that for qualitative results at and near A the O(5) term can be neglected. If ka_4 is chosen to equal one-half of unity, equation (13) becomes

$$V(\theta) \sim b_0 + b_1\beta\theta + b_2\alpha\theta^2 + b_3\beta\theta^3 + \frac{1}{4}\theta^4$$

Further simplifications can be made to eliminate the cubic term by the substitution of

$$x = \theta + b_3\beta$$

and to omit the high order in α and β. This leads to the energy in the form

$$V \sim b_0 + b_1\beta x + b_2\alpha x^2 + b_3\beta x^3 + \frac{1}{4}(x^4 - 4b_3\beta x^3)$$

$$= b_0 + b_1\beta x + b_2\alpha x^2 + \frac{1}{4}x^4$$

Then new variables u_1 and u_2 are defined by

$$u_1 = 2b_2\alpha, \qquad u_2 = b_1\beta$$

The potential function can be described by

$$V(x) = \frac{1}{4}x^4 + \frac{1}{2}u_1x^2 + u_2x + b_0$$

If the origin of the energy values is translated to make $b_0 = 0$, then above equation becomes

$$V(x) = \frac{1}{4}x^4 + \frac{1}{2}u_1x^2 + u_2x$$

which defines the cusp catastrophe.

The critical points of the system are the solutions of dV/dx = 0, or of equation $x^3 + u_1x + u_2 = 0$, and because it is a cubic in x it has either one or three real roots. The number of roots depends on the values of u_1 and u_2,

specifically on the discriminant

$$\Delta = 4u_1^3 + 27u_2^2$$

It is well known that if $\Delta > 0$ there is only one real root; otherwise there are three roots. The roots are different unless $\Delta = 0$, in which case some of them coincide.

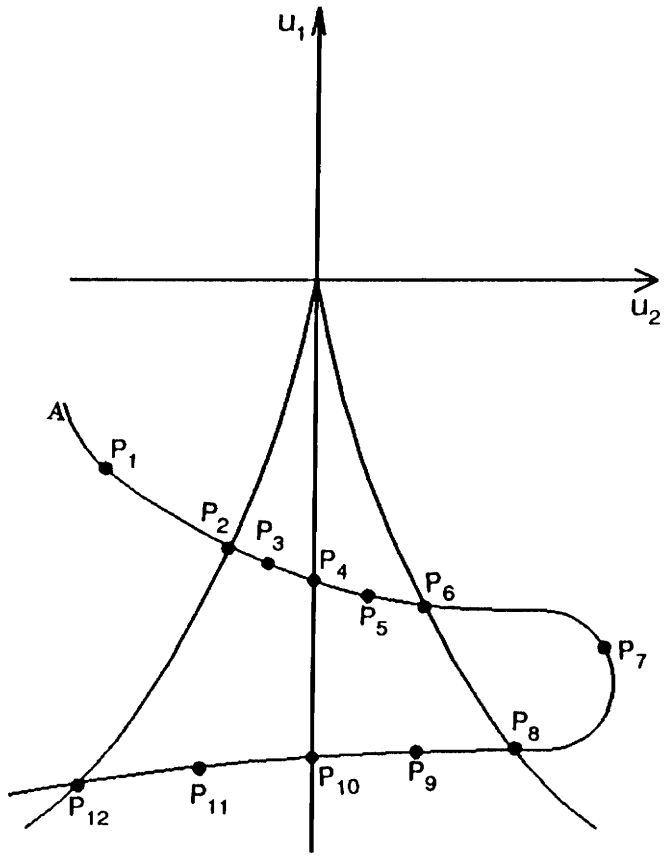
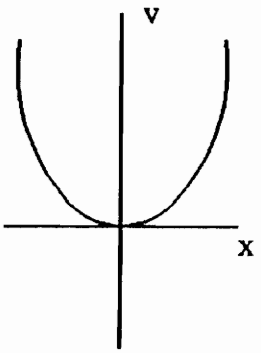


Figure 6 Control manifold

This can be exhibited by a diagram in Figure 6 where $P_0 = P(0,0)$, $P_1 = P(u_{1\text{-}1},u_{2\text{-}1})$, ···, $P_{12} = P(u_{1\text{-}12},u_{2\text{-}12})$ are twelve representative points, and the cusp-

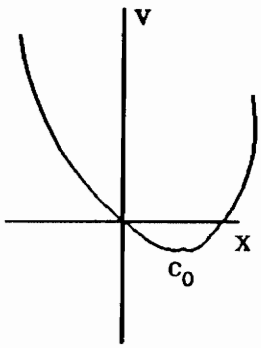
shaped heavy line describes the curve related to equation $\Delta = 0$. The corresponding potential curves to those points are shown in Figure 7.

Starting from any point, say P_3, the corresponding potential is $V(x,P_3)$ as in Figure 7-4. In Figure 7-4, either one of the local minimum $c_0(P_3)$ or $c_1(P_3)$ can be chosen in the beginning. If $c_0(P_3)$ is chosen and P is moving along the curve A in Figure 6, as P approaches towards P_4, according to the delay rule the choice of the local minimum will still be $c_0(P_4)$. In other words, at P_4, there is no sudden change of the location of local minimum. Only when P reaches P_6, does the local minimum $c_0(P)$ disappear and the stable local regime at $c_1(P_6)$ will be chosen. Therefore catastrophic change in state occurs. As $P \rightarrow P_8$, $c_1(P) \rightarrow c_1(P_8)$ which is still a stable minimum. Similarly, as $P \rightarrow P_{10}$, $c_1(P) \rightarrow c_1(P_{10})$, again a stable minimum. But when $P \rightarrow P_{12}$, there is a sudden change from $c_1(P_{12})$ to the stable regime $c_0(P_{12})$. So a jump occurs on leaving the cusp.



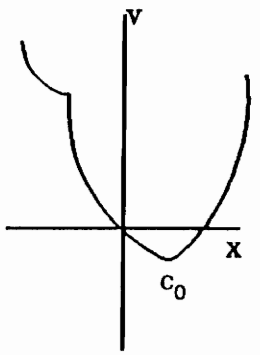
$V(x,P_0)$

7-1



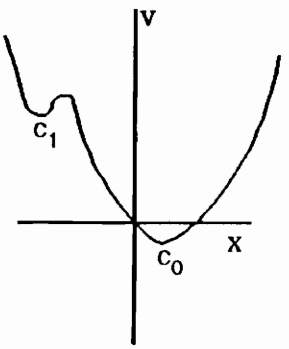
$V(x,P_1)$

7-2



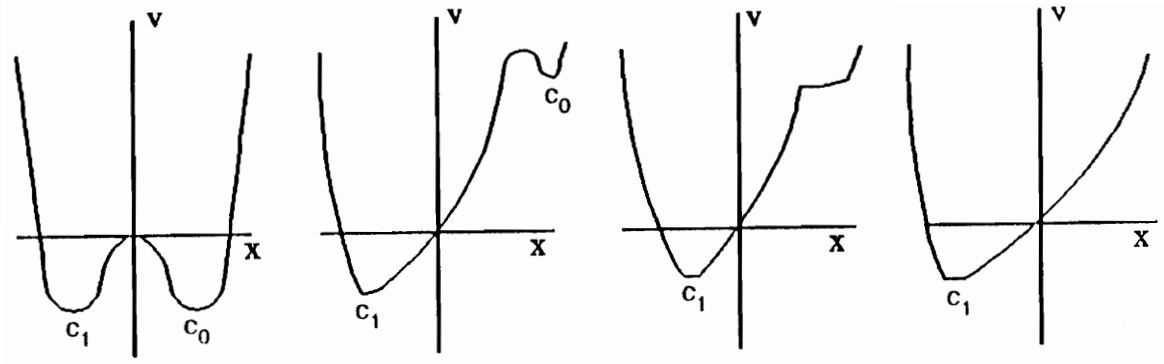
$V(x,P_2)$ or $V(x,P_{12})$

7-3



$V(x,P_3)$ or $V(x,P_{11})$

7-4



$V(x,P_4)$ or $V(x,P_{10})$ $V(x,P_5)$ or $V(x,P_9)$ $V(x,P_6)$ or $V(x,P_8)$ $V(x,P_7)$

7-5 *7-6* *7-7* *7-8*

Figure 7 V(x) for different values of u and v

1.5 The Remaining Catastrophe

If the dimension of the control space is increased or decreased, then this is quite a different kettle of fish. If, for example, control space is 1-dimensional and behavior is 1-dimensional, instead of being folded along curves the catastrophe surface would be degenerated to a parabola curve, and the bifurcation set would be a single point instead of consisting of curves with cusp points in 2-dimensions. This new kind of catastrophe is called fold catastrophe, the simplest of all

catastrophes.

The standard potential equation is

$$V(x) = \frac{1}{3}x^3 + ux$$

where u is the only numerical factor to be introduced to control the whole system. The catastrophe manifold is given by

$$\frac{dV}{dx} = x^2 + u$$

This equation suggests that the manifold is a parabola curve in 2-dimensional space as shown in Figure 8. The second derivative of potential

$$\frac{d^2V}{dx^2} = 2x$$

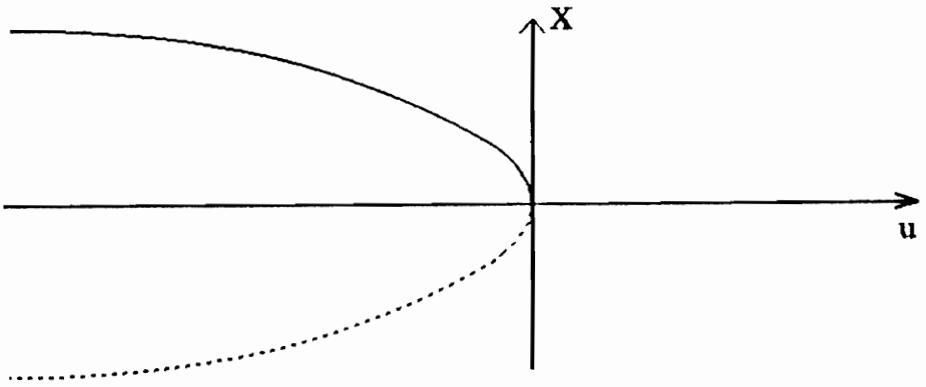


Figure 8 The fold catastrophe

gives the stable and unstable states of the system. Since this is positive for positive values of x, and negative for negative values, this shows that the minima occur for the positive values and the maxima for negative values. The bifurcation set is a single point at origin where $d^2V/dx^2 = 0$. Any pair of values for u and x, or any point of the plane, stands for a single combination of the control factor and behavior. The control manifold, in this case, is simply the horizontal axis. And the bifurcation set is also very simple: it is the single point at origin. It is at this point therefore that jump behavior can be observed. If the system is in a state given by a negative u and on a trajectory in which u is increasing, then as u passes through zero, the stable minimum state disappears and the system will have to take up some other states not accounted for by this diagram. There is little to say about fold catastrophe, since only a few things can happen in such a simple system, all of them obvious.

If the behavior surface is expanded, there are two more kinds of catastrophes. It is impossible to draw the complete picture in each of these cases because this would require more than 3-dimensions. For no more than four control parameters, there are only seven elementary catastrophes describing all possible discontinuities in phenomena which are summarized in Table 1. Each catastrophe is associated with a potential function in which $u_1, \cdots, u_4$ are coefficients for control factors, and x, y are variables for system states. The behavior surface

is determined by $dV/dx = 0$ in single behavior variable cases, or $\partial V/\partial x = \partial V/\partial y = 0$ when there are two first derivatives.

Table 1 Seven Elementary Catastrophes

Name of Catastrophe	Control Dim.	Behavior Dim.	Function V
Fold	1	1	$\frac{1}{3}x^3 + u_1x$
Cusp	2	1	$\frac{1}{4}x^4 + \frac{1}{2}u_1x^2 + u_2x$
Swallowtail	3	1	$\frac{1}{5}x^5 + \frac{1}{3}u_1x^3 + \frac{1}{2}u_2x^2 + u_3x$
Butterfly	4	1	$\frac{1}{6}x^6 + \frac{1}{4}u_1x^4 + \frac{1}{3}u_2x^3 + \frac{1}{2}u_3x^2 + u_4x$
Hyperbolic	3	2	$x^3 + y^3 + u_1x + u_2y + u_3xy$
Elliptic	3	2	$x^3 - xy^2 + u_1x^2 + u_1y^2 + u_2x + u_3y$
Parabolic	4	2	$x^2y + y^4 + u_1x^2 + u_2y^2 + u_3x + u_4y$

Since the fold and cusp catastrophe are the only two catastrophes that can

be drawn in their entirety, and are easy to understand, they are the most frequently used in qualitative applications. This thesis is focused on their characteristics as well as their applications.

2. SYSTEM DYNAMICS

2.1 An Introduction

A model is a summation of a system. It can be represented by a set of rules and relationships that describe something. A model yields information more quickly and at lower cost than working with real systems. In this sense, it is widely accepted in analyzing complex phenomena.

A model can be classified in many ways. It can be static or dynamic, open or closed, etc. A static model represents situations that don't change with time, while a dynamic model describes the relationships with varying time. The former leads to emphasis on discrete symptoms, such as: services don't satisfy the needs of people, etc. And the latter deals with not only the separate elements of a system, but also the interactions between these elements as the system operates through time. Adopting a dynamic model helps people to better understand the information-receiving and decision-making process, continually interacting with a changing reality.

An open model is one whose outputs have no effect on the inputs. There isn't any relationship between the outputs and inputs. Thus, future action is not influenced by past one, and the system is not "aware of" its own performance. On the other hand, a closed model, which is usually called a feedback model, generates the outputs of the variables by the interactions of these variables and uses these outputs to control future actions. One of the characteristics of closed model is that it can exhibit informative behavior without receiving an input from external sources time after time. Normally, the closed model is characterized by a closed loop structure, called a feedback loop, that brings results from previous actions of the system back to control future actions.

System dynamics modeling, developed by Professor Jay Forrester of M.I.T., is a methodology that deals with a dynamic, closed system. The principles and mechanics of system dynamics were first worked out in 1961 in the study of the behavior of industrial systems. Since then, the technique has been used to understand dozens of different problems such as urban development, population growth, environmental deterioration, and so on.

The basic structure of system dynamics is the interacting feedback loops. It is a closed path connecting decisions that control actions, states of the system, and information about the system. Since flow charts offer a convenient way to

represent loop structure, it is no surprise that the causal-loop diagrams are used before development of system equations.

2.2 The Heart of System Dynamics

The importance of structure is it is the guideline in organizing information. A system may at first sight seem meaningless, but it might be classified by fitting it into a limited number of categories after studying its structure. The level, rate, and feedback loop are the concepts of structure to be used to organize the system in system dynamics.

Levels are defined as the accumulations within the system. They, and only they, describe the condition of the system at any particular time. They can be accumulated by the difference between inflows and outflows. The computation of a new level variable is independent of the values of any other levels.

On the contrary, rate variables are quite different. They define the present flows between the levels in the system. In the other words, they measure the changing speed of the levels. The levels corresponded to the state of the system,

while the rates determine the action of the system which leads to that state. In turn the rates of flow are determined by the levels of the system according to the rules defined by the policy makers.

Feedback loops consists of levels and rates. They are the basic building blocks within the system dynamics. They couple the decision, action, condition, and information, with a path returning to the decision. The decision controls the action which alter the system levels which influence the decision. A system may be a single feedback loop or a couple of interconnected feedback loops. No matter how many feedback loops a system has, it always has the following characteristics:

- Be able to describe any statement of causal relationship
- Be able to generate discontinuous changes in decisions when these are needed
- Be simple in mathematical expression

Moreover, the feedback loop structure generates the dynamic behavior of the system. The behavior of the system is associated with the time-sequence relationships between the actions in the system. Time delays exist in each stage of the actions -- in accumulating levels, transporting data, etc. This unusual

attribute determines that the inflow rates need not match the outflow rates at all times.

On the basic idea of feedback loops, the variables in a system can be displayed diagrammatically. Such a flow diagram forms a mean of communicating the nature of the model, and helps to establish the mathematical equations. In addition it helps most when it provides new insights. It shows how levels and rates are interconnected to produce feedback loops and how feedback loops are interlinked to create systems. It discloses what factors enter into each decision function. Since each variable is connected in terms of cause-and-effect interactions, the diagram is called causal diagram.

Since the causal diagram does not reveal what the mathematic functional interactions between each variable, the mathematical equations, representing the feedback system, should be developed simultaneously with the development of the causal diagram. They tell how to generate a new system condition, given the previous system condition. These equations, called DYNAMO equations, are evaluated time by time in a sequence of steps equally spaced in time. In addition to level equations and rate equations, auxiliary equations, table-functions, initial value equations, and constant value equations are used.

3. SYSTEM DYNAMICS AND CATASTROPHE

3.1 Relationship Between System Dynamics and Catastrophe Theory

As it is known to all that system dynamics models can be expressed by a set of differential equations as

$$\frac{dL}{dt} = f(L, C, t)$$

where L is the levels

C is the control parameters

t is the time variable

Most system dynamics models of interest are not gradient systems. That is, the corresponding differential equations cannot be derived from a potential function V as in the form of $f = -\frac{\partial V}{\partial L}(L, C) = -\nabla V$, since in general the process of development of a model does not lead naturally to formulating the net

rate affecting each level as the partial derivative of such a function. Typically, the system dynamics models contain more than four independently determined parameters. Each rate equation, which represents a statement of policy that determines how the available information about levels leads to the current rate, has at least one "overt" or "implicit" parameter to translate the effect of the environment on policy-maker's decision. So in summary the system dynamics models represent complex, non-gradient-type dynamic systems, which are more general than the gradient systems with less than four independent control parameters for which catastrophe theory results are available.

In addition system dynamics models are used to understand the nature of the transition from one equilibrium to another. Most dynamical behavior can only be represented by models so complex that analytical mathematical solutions are impossible. The DYNAMO equations describe how the system changes, and these changes are accumulated step-by-step to disclose the behavior pattern of the system. This simulation method doesn't tell how to go directly to some distant future condition without first computing through all of the intermediate stages. On the other hand, the catastrophe theory models are applied to classify and treat the various equilibrium states of the system. This is a quite different approach of solving the equations that define the behavior of the system to obtain an analytical solution. Such a solution to the equations can be used to express the system

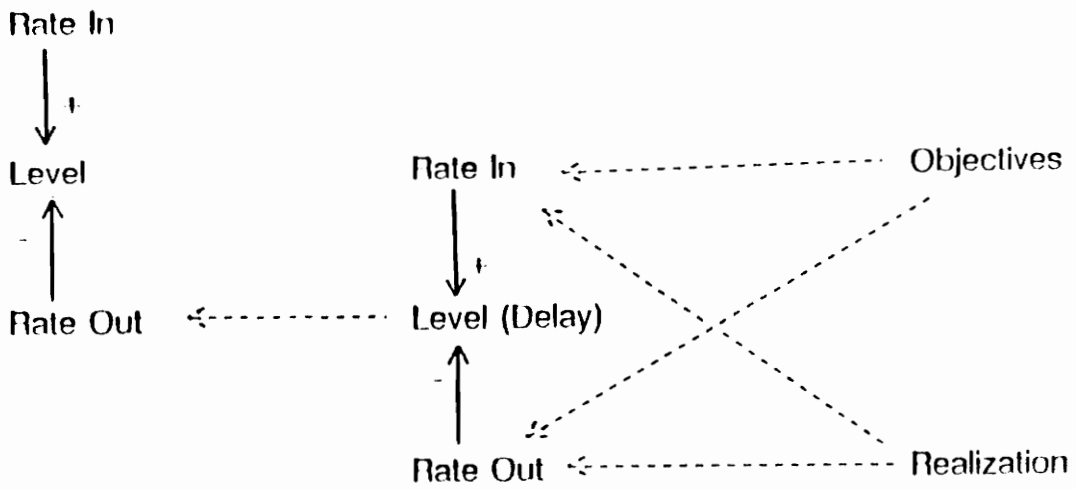
condition in any future time, no matter in terms of a short time interval or a long period of time. For these reasons it is unlikely that the system dynamics models could be reducible to catastrophe theory models.

Although the detailed dynamic system information for which system dynamics is evolved could not be generally derived by catastrophe theory, it is obvious that the spirit in which system dynamics approaches to modeling shares many points in common with the development of catastrophe theory. Especially both are generated to search for a small number of structures corresponding to distinct behavior modes.

As both are used to deal with non-linear systems and to seek to develop fruitful simplifications of a complicated reality, it is no surprise to find catastrophic behavior occurring in system dynamics models. However, the limits of catastrophe theory --- reduction to one or two state variables and less than four control parameters --- as well as the realistic orientation of system dynamics models --- the identifiable definitions of all levels and variables --- constrain the catastrophes occurring in sub-systems of one or two levels and being driven by varying the parameters either exogenously or by relationship between the subsystem and the larger system to which it belongs.

3.2 A System Dynamics Model Equivalent To A Catastrophe Theory Model

As mentioned earlier, one of the characteristics of catastrophe is that the control points within the catastrophe set are determined by delay convention. It is now possible to model subsystems presented in delay under the influence of slow variations in the two factors. The following causal diagram is presented by introduced in a second level variable which represents the delay.



The delay is determined by the use of two variables, which are explained in terms of objectives and achievements. The values taken by the two conflicting

factors and their deviation tend either to increase the value of delay or to decrease its value, or even to cancel it.

4. APPLICATION

4.1 A Preview

In the rest of the thesis, a couple of applications of the method are presented which cover transportation planning and urban studies. These are based on the ideas outlined in the previous three sections. It is useful at this stage to take stock before embarking upon the main section and to review the nature of the types of application.

Catastrophe theory as such was presented as being concerns largely with the topology of surface of possible equilibrium states of a gradient system. The behavior of the system is represented by trajectories on such surface. The path taken by a particular trajectory is determined by the way in which the parameters vary. By using catastrophe theory, it is nearly always assumed that the system stays on the equilibrium surface. This means that the speed of return to equilibrium of the system is considered to be very fast comparing to the changes of the control variables. If there is any disturbance from equilibrium, the system is assumed to move rapidly back to equilibrium.

The mathematics of catastrophe theory is often said to be qualitative rather than quantitative. It is considered that the mathematics is more important for its qualitative insights than for the more traditional style of detailed quantitative modelling. The principle adopted here is that the representation of detailed mechanism of change should be sought and this usually means building a qualitative model. The main use of the catastrophe theory is to alert people to the possibilities of new behavior and to look for parameters with particular roles. This refers to something like the following. Variables are defined to describe the system of interest including an appropriate number of control variables. These control factors can then be used to determine the nature of the system behavior which can occur.

4.2 Application To Urban Transportation Planning

All over the world, cities are growing through migration from rural areas and natural population increase. With the rising of incomes and automobile ownership, the urban residents are more and more willing to live in the low-density suburbs instead of downtown areas. This trend leads to an increase in travel, especially by private automobile. This results in a continuing need for development of

transportation facilities in urban areas.

However, there is also a need to respond to the changes in economic and environment conditions. As both land and energy are becoming scarce sources, the means to achieve the transportation goals may shift away from emphasis on increasing urban mobility toward seeking optimization of the efficiency of the whole network. This requires the planner to consider the transportation capacity, present and projected land use, and socio-economic development at the same time.

Many efforts at building models for use in the urban planning process so far have been made. Among all of these techniques, system dynamics is a suitable, efficient tool to display the models comprehensively. It is dynamic, concerned with system behavior as fully as possible, and based on causal relationships. Multiple goals are involved. The causal diagram shown in Figure 9 depicts the interrelationships among transportation facilities, industrial development as well as land use. The background of the model is one of population mobility, attracted or repulsed by favorable or unfavorable local transportation conditions. The model matches the industrial demand by the supply of labor available. The structure of the model also relates the development of the regional industry with the highway construction fund which is a major source to determine the development of transportation facilities. Since the increase of the

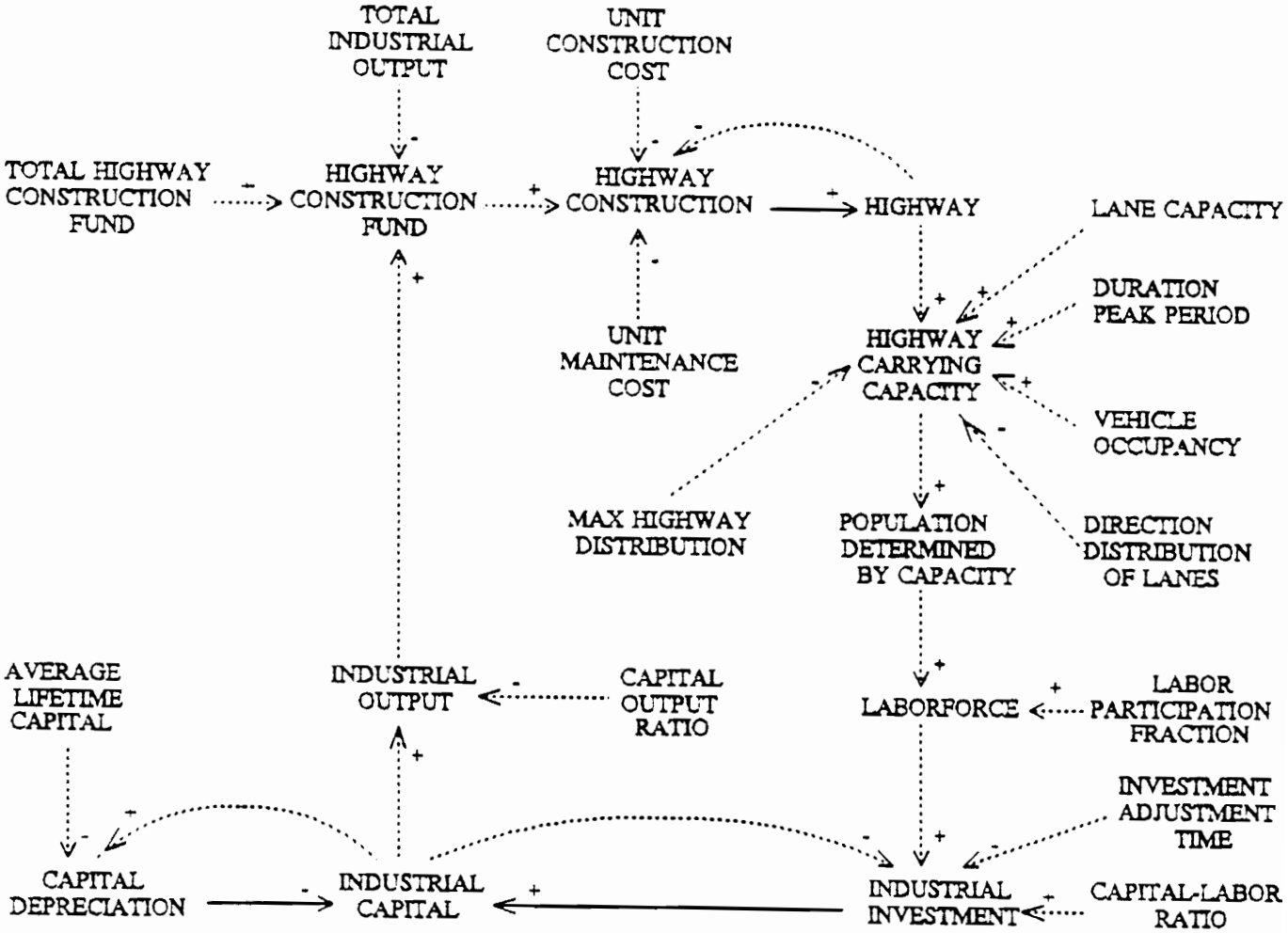


Figure 9 Causal diagram

```
L     IC.K=IC.J+(DT)(II.JK-CD.JK)
N     IC=ICN
C     ICN=?
NOTE IC-INDUSTRIAL CAPITAL ($)
R     CD.KL=IC.K/ALC
NOTE CD-CAPITAL DEPRECIATION ($/YR)
C     ALC=30
NOTE ALC-AVERAGE LIFETIME CAPITAL (YR)
A     IO.K=IC.K*COR
NOTE IO-INDUSTRIAL OUTPUT ($/YR)
C     COR=2
NOTE COR-CAPITAL OUTPUT RATIO (1/YR)
A     HCF.K=IO.K*THCF/TIO
NOTE HCF-HIGHWAY CONSTRUCTION FUND ($/YR)
C     THCF=6E13
NOTE THCF-TOTAL HIGHWAY CONSTRUCTION FUND ($/YR)
C     TIO=8E11
NOTE TIO-TOTAL INDUSTRIAL OUTPUT ($/YR)
R     HC.KL=(HCF.K-H.K*UMC)/UCC
NOTE HC-HIGHWAY CONSTRUCTION (MILE-LANE/YR)
C     UMC=20000
NOTE UMC-UNIT MAINTENANCE COST ($/MILE-LANE)
C     UCC=200000
NOTE UCC-UNIT CONSTRUCTION COST ($/MILE-LANE)
L     H.K=H.J+(DT)(HC.JK)
NOTE H-HIGHWAY (MILE-LANE)
A     HCC.K=H.K*LC*DPP*VO/(MHD*DDL)
NOTE HCC-HIGHWAY CARRYING CAPACITY (PERSONS)
C     LC=2000
NOTE LC-LANE CAPACITY (VEH/LANE-HR)
C     DPP=3
NOTE DPP-DURATION PEAK PERIOD (HR)
C     VO=1.4
NOTE VO-VEHICLE OCCUPANCY (PERSONS/VEH)
C     MHD=30
NOTE MHD MAX HIGHWAY DISTRIBUTION (MILE)
C     DDL=0.5
NOTE DDL-DIRECTION DISTRIBUTION OF LANES (DIM)
A     PAHCC.K=TABLE(PAHCCT,HCC.K,?,?,?)
NOTE PAHCC-POPULATION ATTRACTED BY HIGHWAY CARRYING CAPACITY (PERSONS)
T     PAHCCT=?/?/?/?/?
A     L.K=PAHCC.K*LPF
NOTE L-LABORFORCE (PERSONS)
C     LPF=0.4
NOTE LPF-LABOR PARTICIPATION FRACTION (DIM)
R     II.KL=(L.K*CLR-IC.K)/IAT
NOTE II-INDUSTRIAL INVESTMENT ($/YR)
C     CLR=50000
NOTE CLR-CAPITAL-LABOR RATIO ($/PERSON)
C     IAT=5
NOTE IAT-INVESTMENT ADJUSTMENT TIME (YR)
```

Figure 10 DYNAMO equations

transportation capacity enhances the attractiveness of the region, and therefore attracts the socio-economic activities, the feedback loop is closed. However, there will doubtlessly be conflicts between different parts of the community. The growth in urban activities such as residential, industrial, and commercial development, recreational facilities and public open space depends on the land use. The objective of the model, therefore, is to develop orderly programs under which transportation system can be fully developed as well as the local socio-economic activities.

Because of their simple and highly aggregated structure, the model used in general is based on simplifying assumptions. One of these assumptions is the linear allocation of transportation funds according to the regional industrial output. The more the industrial output, the more the transportation construction fund used in this region. The other one assumes that the immigrated population grows exponentially in early stage but is then limited by an upper bound. This is a common feature of urban systems while the bound is created by land limits, pollution, etc.

The DYNAMO equations for the model described in Figure 9 are given in Figure 10. In order to investigate the trajectory, the differential equations of the model might be examined.

The resulting model can be expressed by a pair of ordinary differential equations which express the rate of change of each level variable in terms of the current values of those variables:

$$\frac{dIC_t}{dt} = \frac{CLR \times LPF}{IAT} \times F\left(\frac{LC \times DPP \times VO}{MHD \times DDL} \times H_t\right) - \frac{IC_t}{ALC} - \frac{IC_t}{IAT} \qquad (14)$$

and

$$\frac{dH_t}{dt} = \frac{THCF \times COR}{UCC \times TIO} \times IC_t - \frac{H_t}{UCC \times UMC} \qquad (15)$$

where $F\left(\frac{LC \times DPP \times VO}{MHD \times DDL} \times H_t\right)$ describes the table function of relationship between population and transportation facilities in the previous DYNAMO equations.

These equations are usually non-linear, since the right-hand sides involve more complicated functions than the first powers of H_t and IC_t, and therefore do not involve simple, explicit solutions. But they can be solved numerically, and H(t) and IC(t) can be calculated by simulation once H_0 and IC_0, their values at $t = 0$, as well as other policy parameters are given. Very often the values of H_t and IC_t approach constant, or steady-state values.

It is clear from equation (14) and (15) that the system is behaving in a manner that is not easily accounted for by a model based on gradient system equations. To illustrate the possibilities of catastrophe in this model, it is necessary to explore the equilibrium points and trajectories on phase space diagrams. The equilibrium points of the system are the solutions of

$$aIC_t - bH_t = 0 \tag{16}$$

which intersect with those of

$$cF(H_t) - dIC_t = 0 \tag{17}$$

where $a = \frac{THCF \times COR}{UCC \times TIO}$,

$b = \frac{1}{UCC \times UMC}$,

$c = \frac{CLR \times LPF}{IAT}$,

and $d = \frac{1}{ALC} + \frac{1}{IAT}$.

Equation (16) is a typical line equation. However, the trajectory of equation (17) depends on the function $F(H_t)$. As noted in the second assumption, the function $F(H_t)$ narrates the connection between the immigration and highway development,

and it can be expressed by one of the curves in Figure 11 on the next page, or equation

$$\frac{dF(H_t)}{dH_t} = \alpha [\beta - F(H_t)] [F(H_t)]^n \qquad (18)$$

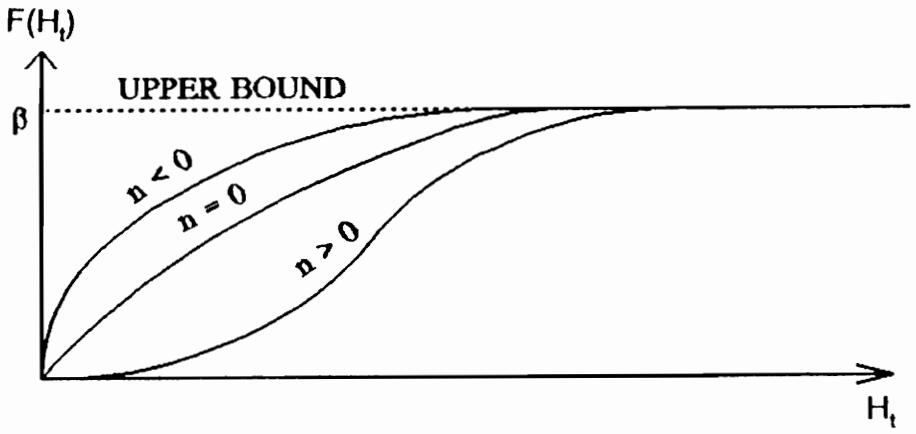


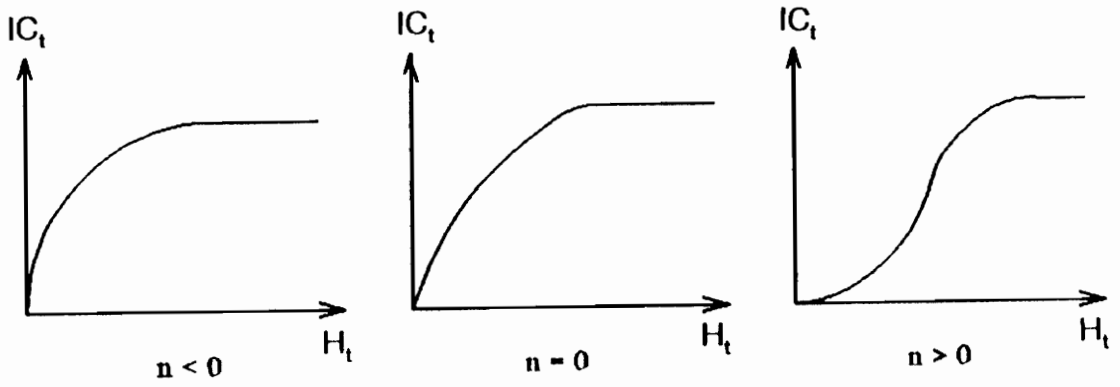
Figure 11 Population - highway capacity curves

It can be easily seen from equation (18) that β is the upper bound of growth and so it is sometimes known as the carrying capacity of the system. An important case which turns up frequently is $n = 0$. In this case, $dF(H_t)/dH_t$ is finite and non-zero at the origin and so the immigration population there grows faster than that for the logistic curve. For $n < 0$, the gradient of $F(H_t)$ at the origin becomes infinite while for $n > 0$, the gradient is always zero as in the logistic case.

The inflection is more accentuated as n increases.

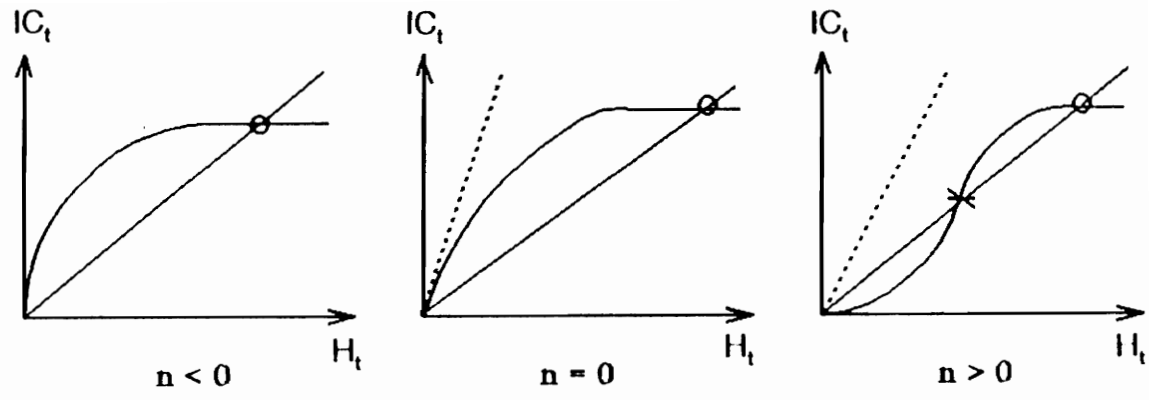
Now it is possible to plot IC_t as a function of H_t. Since the IC_t is proportional to $F(H_t)$ from equation (17) the IC_t - H_t curve is obviously the same shape as the $F(H_t)$ - H_t. For $n < 0$, $n = 0$, and $n > 0$, the plots are in Figure 12.

If the line (16) is added to the curves of Figure 12, the steady-state of the system is simply determined by the intersections of these two curves. The basis for this is Figure 13. The two kinds of intersection of curve and line are shown in Figure 14.



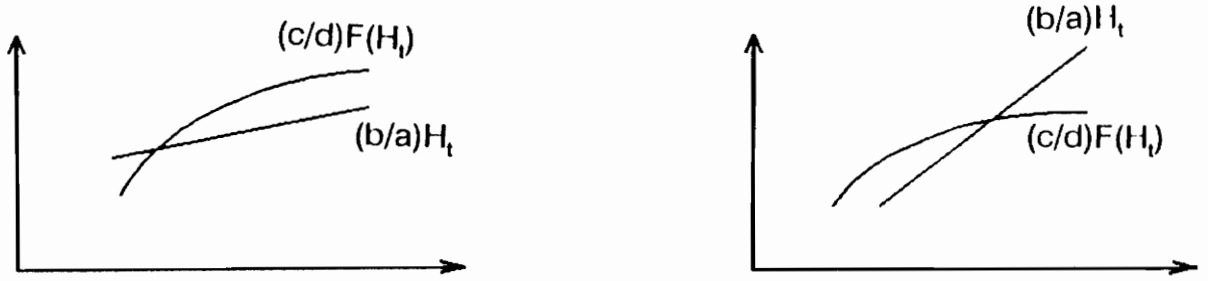
12-1 *12-2* *12-3*

Figure 12 Industrial capital - highway capacity curves



13-1 *13-2* *13-3*

Figure 13 Illustration of equilibriums



14-1 *14-2*

Figure 14 Stability considerations

When the curve is above the line, (c/d)F(H$_t$) - (b/a)H$_t$ is positive; and vise

versa. This means that in case 14-1, the industrial development affected by population is greater than that determines the highway construction, and it leads to further increase in the highway construction to meet the demand of the economic development. Hence the point is unstable. The reverse argument applies to case 14-2, which is a stable point. This argument is applied to the various intersections on Figure 13: the stable points are circled and the unstable point marked with a cross.

The next step to note is that the line does not always intersect the curve: two cases are shown on each of the plots in Figure 13 -- the dashed line obviously being the non-intersecting case.

To illustrate the idea of catastrophe, it is focused on the $n > 0$ case, and on changing values of the parameter b/a or θ, if θ is represented the slope of line. The $IC_t = \theta H_t$ line is plotted with three different θ values in Figure 15. In case (a), there are two possible stable points; in case (c), there is only one, the origin; and case (b), the slope of straight line θ, is critical. The situation changes from one where highway construction is possible (low value of θ) to one where it is not (high value of θ). If initially, the value of θ is reciprocal to that in case (a) which intersects the curve at A and B, unstable and stable equilibrium points respectively, and H_t is at H_1, then H_t tends to grow to H_B where the system reaches the steady-

state. If for some reasons not relating to this system, the value of θ starts to increase, the straight line in case (a) will rotate to one in case (c) passing through case (b). During this transition, the equilibrium point of H will shift to H^{crit}, and then suddenly to origin H_0 as the straight line no longer contacts with the curve. This sudden change in the equilibrium suggests that catastrophe theory might provide

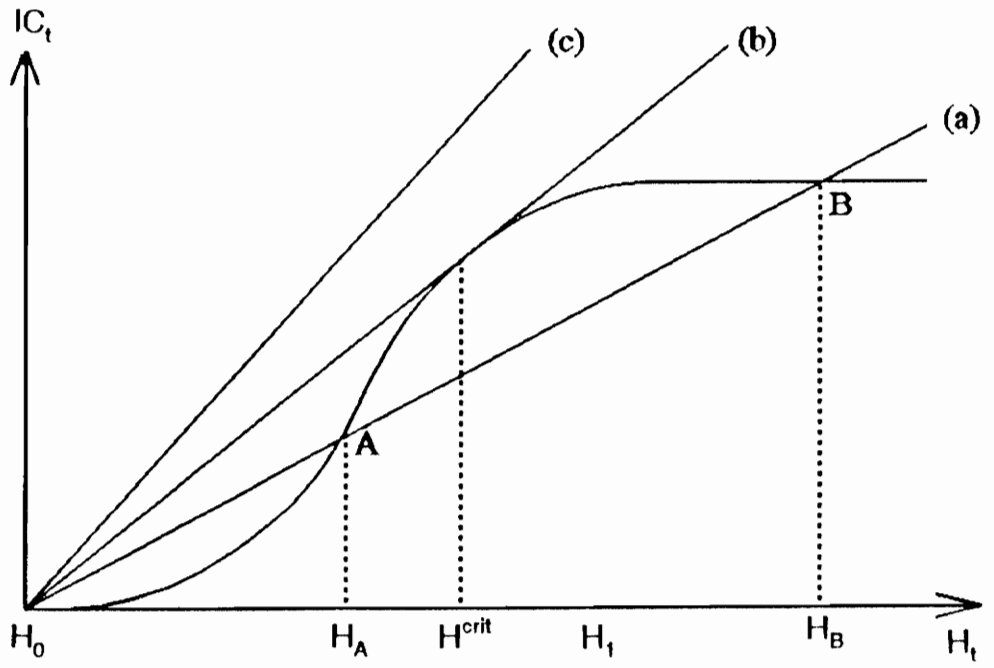


Figure 15 Illustration of different equilibriums

a suitable tool for analysis. The idea is to choose a suitable catastrophe and to find the observations to fit. Since the catastrophe model with only one control variable is the fold, it is not surprising that the fold catastrophe is selected to

analyze the model. The state variable is of course H_t at steady-state, since it is in this quality that the discontinuities were observed. An obvious choice for the control variable is the parameter b/a, or the slope of the straight line θ.

With this choice of variables, the following figure (Figure 16) becomes a drawing of the equilibrium values of H_t against a varying θ value. This curve is immediately reminiscent of the fold catastrophe, but with zero states added.

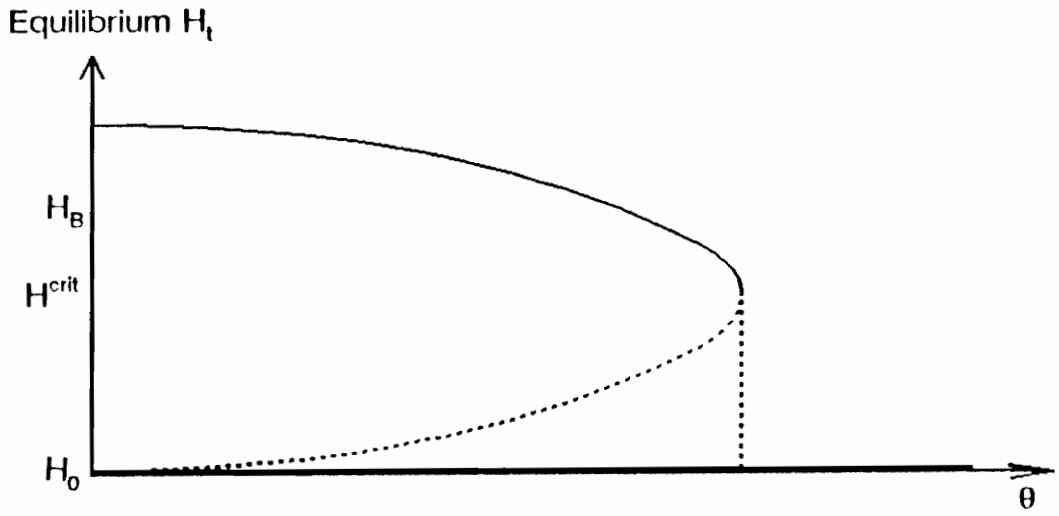


Figure 16 The fold catastrophe of the system

The argument above has been long and complicated. It might be useful therefore to summarize the conclusions and to see what contribution catastrophe theory begins to make towards the evolution of urban transportation structure.

The main general result to emerge is that there is a surface in parameter

space on one side of which development is possible, on the other side not. This was seen clearly with the role of b/a parameters. It was also seen that $n = 0$ is an important critical value as shown in Figure 12: for $n < 0$, H_t is always in the development state; for $n > 0$, this is not the case.

When an equilibrium H_t switch from the non-development state to the development state this will be recorded as an "abrupt jump" in that particular H^{crit} value. Any changes in the policy parameters, THCF, COR, TIO, and UMC which determine the value of a and b will cause the equilibrium of system to shift to another state.

For example, the steady-state of H_t changes from H_b to H^{crit}, as the slope of straight line continuously and smoothly increases, as shown in Figure 16. But any further increase in the slope of line, b/a or θ, will produce a suddenly fall of equilibrium of the system form H^{crit} to H_0. Since the gap between H^{crit} and H_0 is much larger than any gap it has experienced, thus the change of the equilibrium in H_t is rapid and abrupt. However, the real level of H_t does not shift so rapidly. It evolves smoothly and continuously toward the accessible steady-state.

This concludes that the catastrophe phenomenon is embedded in this model. The system equilibrium is determined by the policy parameters and the

shape of assumed curve. Any excessive increase in the rate of TIO / (UMC * THCF * COR) or the value of b/a will lead to a catastrophic jump of the steady-state of the system. This causes the deterioration of the transportation facilities as well as the regional economic conditions.

4.3 Application To Modal Split

In order to select a suitable value of vehicle occupancy in the previous model, the ridership of different kinds of transportation mode needs to be predicted. In other words, the modal split analysis should be executed.

The modal split is considered an issue of choice among competing alternatives in urban areas, the basic split analyzed is between the use of mass transit and private automobile, although submodal split model is sometimes developed to describe the relative use of different forms of mass transit. The purpose of modal split analysis is to predict the proportional use of two or more modes of travel. The modal split model uses generalized travel cost differences between the trip by auto and the trip by transit, and then compute the probability that a user with specific characteristics and trip purpose would use transit. These

are generally formulated by summarizing trips for specific types of travelers. From the viewpoint of transportation engineers, the transit ridership, if taken as a utility function, can often sustain indefinite exponential growth in its early stage but is then limited by an upper bound which is created by human factors.

The equation which represents this process can be expressed by

$$TR = R\frac{e^{-z^b}}{e^{-z^b} + e^{-z^a}} \tag{19}$$

where z^b and z^a describe the transit and auto utility functions respectively, TR is the transit ridership, and R means total ridership.

Due to the inseparable connection between transportation facilities and utility function, the utility function can be described as a function of travel cost, parking cost as well as travel time, waiting time which are converted to dollars, or mathematically expressed as

$$z = \alpha t + \beta$$

where t describes the travel time and waiting time, z is the utility function, and α and β are the constants that can be evaluated from regression. In order to simplify the problem, the auto utility is assumed as a constant q, only transit travel

time is under consideration. Now, equation (19) can be represented as

$$TR = R\frac{e^{-\alpha t-\beta}}{q + e^{-\alpha t-\beta}} \tag{20}$$

where t is the headway of transit.

Since the transit headway is the reverse of its frequency, the transit ridership can be described by using frequency shown in equation (21) instead of headway:

$$TR = R\frac{e^{-\frac{\alpha}{f}-\beta}}{q + e^{-\frac{\alpha}{f}-\beta}} \tag{21}$$

The above equation is widely used to estimate the number of transit ridership in the process of urban transportation planning process. The planner will adjust the mass transit system according to the expected ridership. This equation may be viewed as a transit ridership demand equation.

If it is assumed that the managers of the transit company balance the ridership and transit frequency, and that γ and δ are the suitable coefficients to describe this condition, then such a balancing state can be written as

$$TR = R(\gamma f - \delta) \tag{22}$$

Substituting equation (22) into (21) gives

$$R(\gamma f - \delta) = R\frac{e^{-\frac{\alpha}{f}-\beta}}{q + e^{-\frac{\alpha}{f}-\beta}} \tag{23}$$

The two sides of equation (23) represent different ways of computing the transit ridership: the left hand side is obtained from the suppliers' behavior; and the right hand side represents the consumers' behavior. If both sides are defined by separate functions, calling $TR^{(1)}$ and $TR^{(2)}$ respectively, then

$$TR^{(1)} = R(\gamma f - \delta) \tag{24}$$

and

$$TR^{(2)} = R\frac{e^{-\frac{\alpha}{f}-\beta}}{q + e^{-\frac{\alpha}{f}-\beta}} \tag{25}$$

These can also be usefully called the "supply" and "demand" curves respectively. If $TR^{(1)}$ and $TR^{(2)}$ are plotted separately against f, then the steady-state points are

just the intersections of these two curves. This is simple for equation (24) since $TR^{(1)}$ is a straight line if γ and δ are assumed to be constants. So the main task is to investigate $TR^{(2)}$ as a function of f.

The initial step is to look at the first derivative of $TR^{(2)}$ with respect of f

$$\frac{dTR^{(2)}}{df} = \alpha^2 qR \frac{e^{-\frac{\alpha}{f}-\beta}}{f^2(q + e^{-\frac{\alpha}{f}-\beta})^2}$$

(1) As $f \to 0$, $TR^{(2)} \to 0$

$$\frac{dTR^{(2)}}{df} \to 0$$

(2) As $f \to \infty$, $TR^{(2)} \to R\frac{e^{-\beta}}{q + e^{-\beta}}$

$$\frac{dTR^{(2)}}{df} \to 0$$

This information now allows of presenting the form of the $TR^{(2)}$ - f curve in

the following figure (Figure 17).

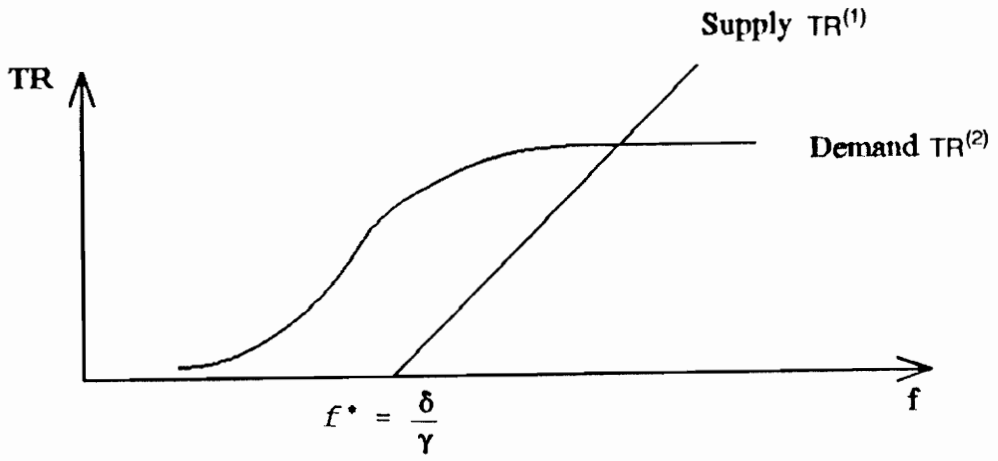


Figure 17 Supply - demand curve

It is now possible to move to the heart of the argument: by adding the lines $TR^{(1)} = R(\gamma f - \delta)$ to the curve of Figure 17, and their intersection gives the equilibrium.

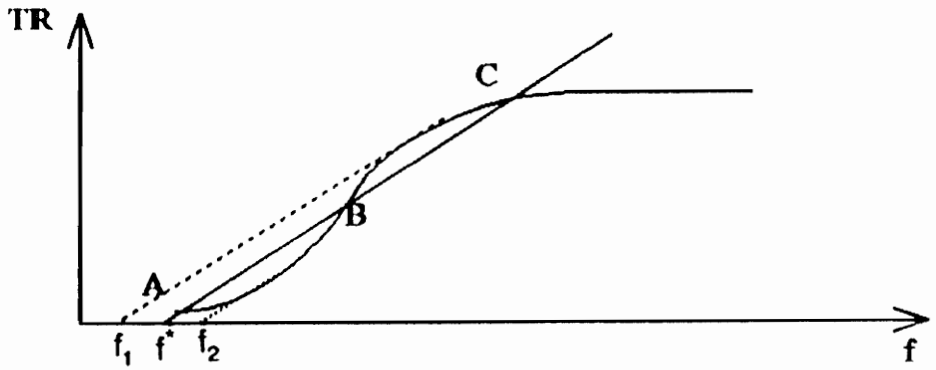
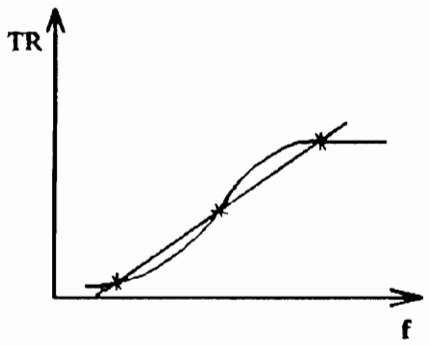
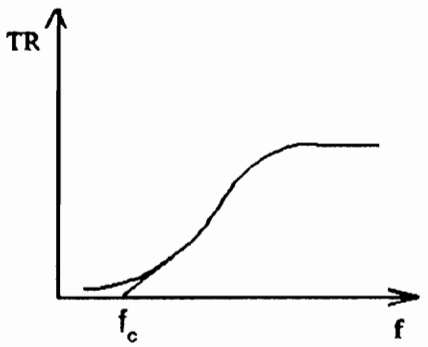


Figure 18 Supply - demand curve

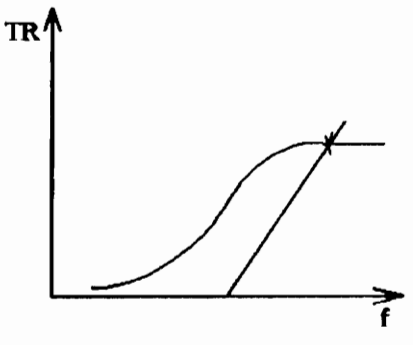
But what will take place if the slope of straight the supply line is so small that the diagram looks like that in Figure 18? If the slope of supply line is smaller than the greatest slope of the transit ridership demand curve, then this figure shows that there will always be two frequencies f_1 and f_2 existing such that if $f^* = \delta/\gamma$ is between f_1 and f_2 there will be three intersections as A, B, and C. This means that there are three possible equilibrium values of transit ridership for any such value of f^*. Clearly the values of f_1 and f_2 depend on the slope of the supply line; if this is less than the maximum slope of the transit demand curve, as in Figure 19-1, then the transit ridership at steady-state expressed as a function of f^* can be plotted by changing f^*. Figure 19-1* shows this curve that folds back on itself between f_1 and f_2. If the line is steep, as in Figure 19-3, then the equilibrium ridership curve as a function of f^* which is shown in Figure 19-3* is very like the curve of transit demand ridership. There is a critical value of γ where it is just equal to the greatest slope of the demand curve and the interval of f_1 and f_2



19-1



19-2



19-3

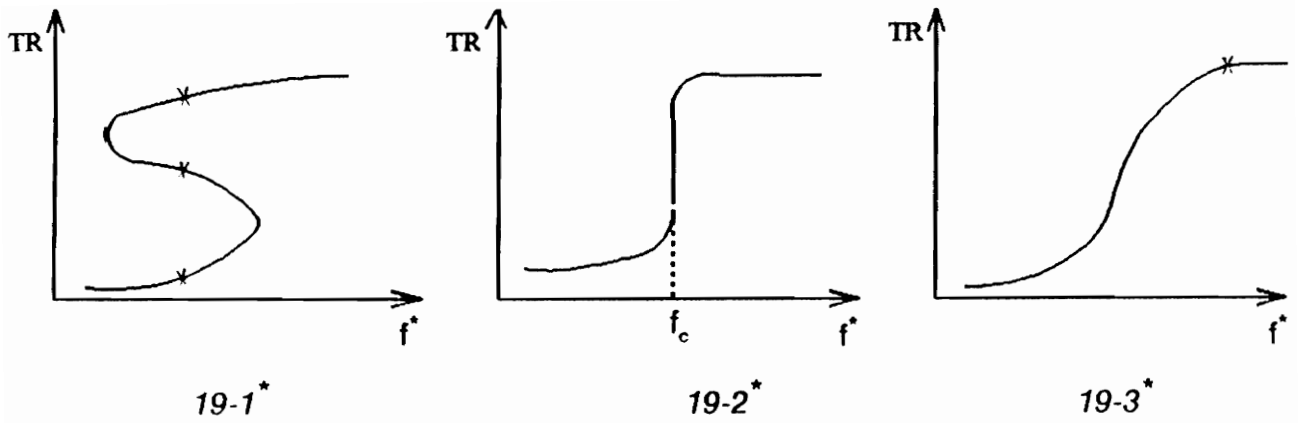


Figure 19 Demand curves with varying supply lines, and equilibrium ridership curves against frequency

a function of f^* gives a unique equilibrium but has a vertical tangent at f_c.

This can be represented very neatly by one of Thom's elementary catastrophes, the cusp. The cusp catastrophe is shown in Figure 20. The control variables in the cusp catastrophe are the transit frequency f^*, and γ; the behavior variable is the transit ridership at steady-state. The behavior surface can be seen to have two regions, a high plateau where the demand is relatively high and a low plain where is very low. For high values of γ (at the back of the diagram) the transition from low ridership to high one is a smooth one. But for a low γ, the transition is catastrophic. Thus, if the transit frequency is increased from such a point as A, the transit ridership at steady-state remains low until C is reached;

when sudden jump to D takes place. Further increase of frequency to E increases the equilibrium values of transit ridership only slightly; but when f^* is decreased again, the jump occurs at F instead of D, when the ridership suddenly drops to B.

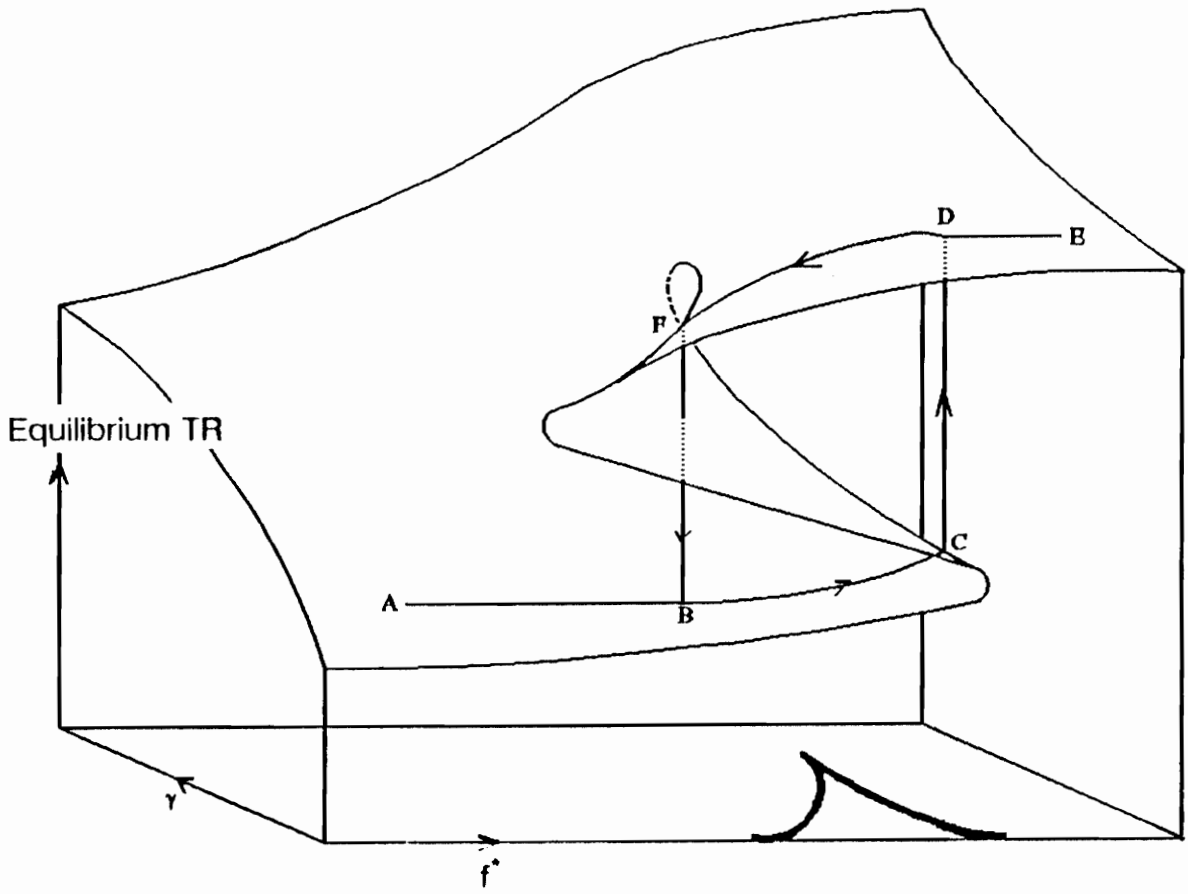


Figure 20 The cusp catastrophe of the system

There is no simple way of getting any point represented by the underfold of the catastrophe surface, and it comes as no surprise therefore if it corresponds to unstable steady-state. An unstable steady state is one that is theoretically

possible but cannot be realized in practice, since the slightest deviation from it would grow until the system finished in a quite remote state. By contrast, a steady state is locally stable if any sufficiently small deviation dies away and the system returns to the original state from which it was displaced. The unique steady state in Figure 19-3 is stable; in fact it is globally stable since, no matter how large a displacement is made, the state always returns to the critical point.

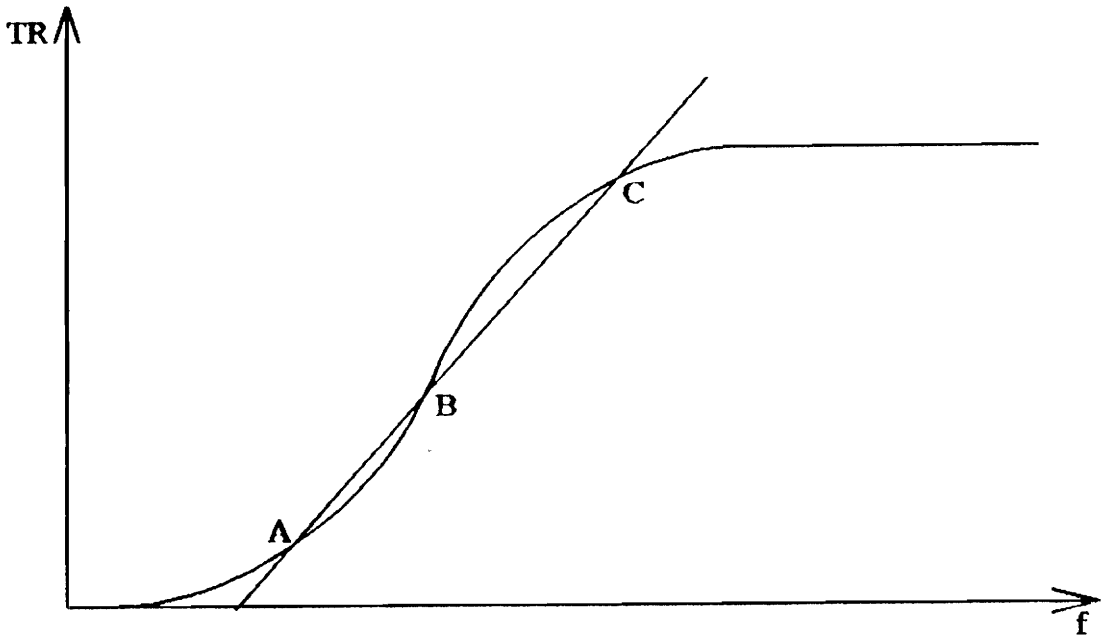


Figure 21 Illustration of stabilities

The conventional argument to investigate the stability of the intersections is demonstrated in Figure 21. The intermediate steady state B is such that if the frequency is increased slightly, the demand exceeds the supply, a situation that obviously leads to even higher transit frequency. Conversely, if the frequency falls,

the transit supply is greater than the transit demand and the frequency continues to fall. This means the point is unstable. The reverse argument applies to point A and C, which are stable points.

Until now the description of catastrophe theory is abstract. In any case, in applied work while it is valuable to know the possible existence of critical points, it is often even more valuable to quantify them. An engineer's ambition is always to seek to carry this through to building a quantitative model.

One of the major challenges to traffic engineers and transportation planners is how to ensure that highways have operationally and economically efficient services which enhance the ridership from suburbs to CBD in the morning peak period, and vice versa in the afternoon. If it is assumed that a suburb is connected with CBD by three-lane in each direction highway which can carry 1800 vehicles per-lane per-hour, and that the ridership from the suburb in the morning peak period is 7500 person-trips per-hour. Then the blockage of the roadway is unavoidable. To lessen the bottleneck phenomenon, providing commuter buses is a good strategy. Moreover, demand metering for auto is used to make more efficient use of roadway. The maximum metering rate is 1800 vph to make sure that the volume in downstream does not exceed the capacity. This is plotted in Figure 22.

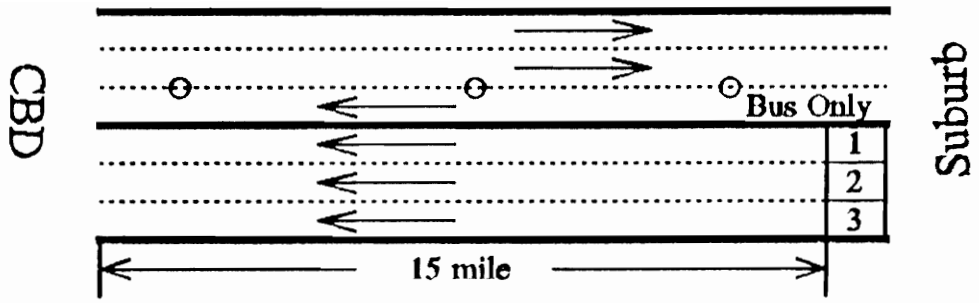


Figure 22 Traffic conditions in morning peak period

It is worth noting that the commuter buses are given preferential treatment. The purpose is to make them immune to the congestion that occurs at peak hour, and therefore to keep the schedule on time. Since the traffic between CBD and suburb is usually heavily directionally unbalanced at peak hour, a control flow bus lane is used to save the investment. The lane is marked by overhead lane signals and separated by short rubber posts. Modal split between auto and transit conforms to equation (26).

$$P_T = \frac{0.4e^{-z_T}}{e^{-z_A} + e^{-z_T}} \qquad (26)$$

where P_T = percentage of transit ridership

$z_A = 1 + 6T_A$

$$z_T = 0.50 + 6T_A + 8T_w$$

The transit service travels at 30 MPH on the average. The waiting time T_w and its frequency comply with the following relationship:

$$T_w = \frac{1}{2} \times Headway = \frac{1}{2}\frac{1}{f}$$

In peak hours, the auto volume on the freeway is often reaching the capacity because of the toll facilities. The excessive auto volume has to stay in the upstream. Thus the travel time for auto in the downstream is 0.50 hour if 60 MPH freeflow speed is assumed. Now equation (26) can be simplified by substituting these numerical values:

$$P_T = \frac{0.4e^{-3.5-\frac{4}{f}}}{e^{-4} + e^{-3.5-\frac{4}{f}}} \qquad (27)$$

If the transit ridership supplied is assumed as

$$R_T = \alpha f \qquad (28)$$

where α is the carrying capacity of the commuter buses, then it is possible to plot equation (27) and (28) in a same diagram. This is done in Figure 23.

The critical point at which the line is tangent with the curve is f = 3.0 or R_T = 909. So the critical α is 303. As noted earlier, this figure is a typical example

of fold catastrophe. The equilibrium points of transit ridership as a function of α are plotted in Figure 24, where the dotted curve expresses the unstable equilibrium points. As the α is greater than α^{crit}, there is only one equilibrium point, the origin. This means the real transit ridership is always lower than that expected by the transit company. In order to cut down the expenses, the company can either reduce the frequency of the commuter buses or decrease the carrying capacity to keep the supply exactly the same as demand. The former tactic is infeasible since any decrease in frequency will further reduce the demand as shown in case (c) in Figure 23. But the latter one is a good strategy. It switches case (c) to case (a) in Figure 23. If the transit company can increase the frequency to 10 per-hour, then the total transit ridership at equilibrium will be 1575, and the auto ridership 5925. It can largely cut down the auto ridership, and lessen the burden of the freeway. From Figure 23 and 24, it can be seen that the key factor influencing the transit ridership is the transit frequency not the capacity. Any excessive increase in capacity with the decrease in frequency will lead to a breakdown of the transit facility. But a suitable frequency will alleviate the chaos of the upstream traffic.

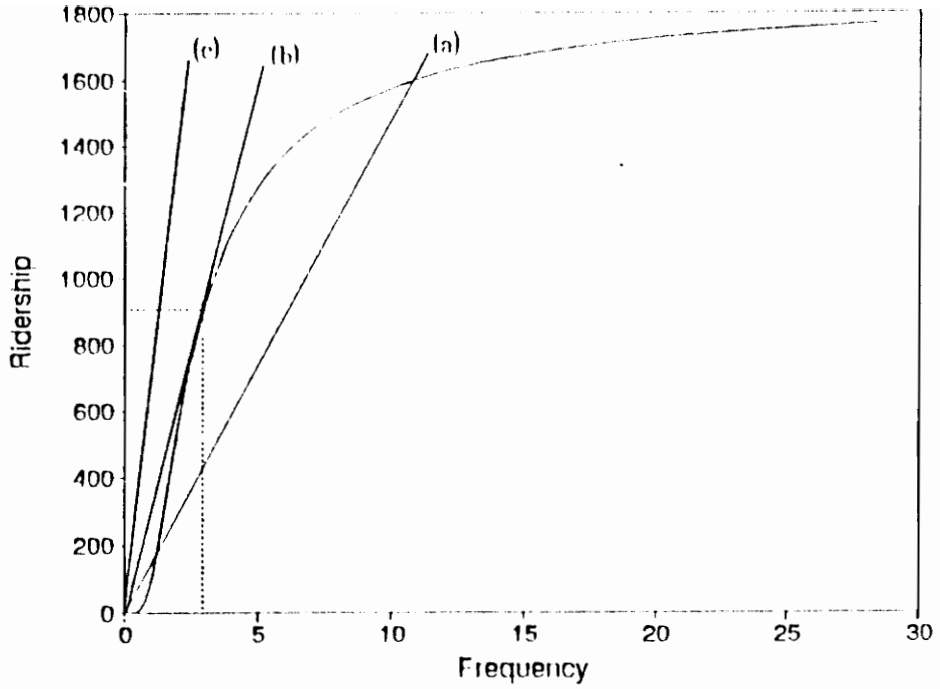


Figure 23 Transit ridership demand curve with varying supply lines

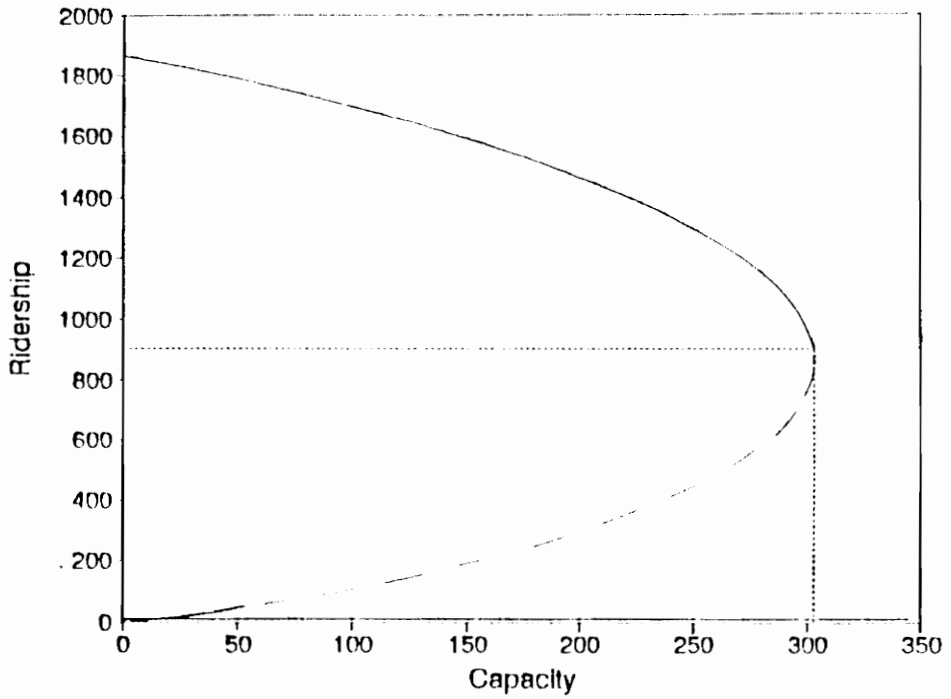


Figure 24 Illustration of the fold catastrophe in transit ridership

5. CONCLUSION

At its present stage, catastrophe theory only deals with systems in equilibrium. It has little to add to an understanding of a period of transition from one steady-state to another. As such the utility of catastrophe theory is limited to equilibrium-seeking systems.

The question that is often asked is whether these phenomena of mathematical models, such as fold and cusp, have any practical importance. In many cases this interesting action takes place within a very narrow range of parameter values, and it is argued that this means that this abstruse behavior can be ignored. But it is possible that in some cases the policy makers are called upon to operate systems under this condition. Thus by analysis of the systems whose parameters are determined by some policy rules which provide the necessary explanation for the appearance of catastrophe, the policy makers can evaluate the effects of a wide variety of policies with some confidence in their conclusions. From a long-term point of view it is a theory, by which new concepts can arise and new understanding can be achieved.

There is the possibility that this method can be applied to alternative models

of the same systems studied here, or the different systems. For further research, it is suggested to run the model by using as much real data as possible, supplemented by realistic hypothesis. Only through numerical experiments it is possible to get new insights for both theoretical and practical purposes. And this adds some flesh to the theoretical bones.

6. REFERENCES

1. Drew, D.R., Graphic Aid Summary For Applied Systems Engineering, Kinko, Blacksburg, VA, 1990.
2. Drew, D.R., System Dynamics: Modeling And Applications, Kinko, Blacksburg, VA, 1990.
3. Drew, D.R., Traffic Flow Theory And Control, McGraw-Hill Book Co., New York, 1968.
4. Forrester, J.W., Industrial Dynamics, M.I.T Press, Cambridge, MA, 1961.
5. Forrester, J.W., Principles of Systems, Wright-Allen Press, Inc., Cambridge, MA, 1968.
6. Forrester, J.W., Urban Dynamics, Wright-Allen Press, Inc., Cambridge, MA, 1968.
7. Forrester, J.W., World Dynamics, M.I.T. Press, Cambridge, MA, 1971.
8. Goodman, M.R., Study Notes In System Dynamics, Wright-Allen Press, Inc., Cambridge, MA, 1974.
9. Liu, J.W., "Future Capital Investment Scenarios In Hong Kong: A Catastrophe Theory Application", Modeling And Simulation, vol.15, 69-74.
10. Martin, M., "System Dynamics And The Catastrophe Theory", System Dynamics And The Analysis of Change, ed. by B.E. Paulre, North-Holland,

1981.

11. Poston, T. & Stewart I., Catastrophe Theory And Its Applications, Pitman, London, 1978.

12. Rahn, R.J., "System Dynamics And Catastrophe Theory", System Dynamics And The Analysis of Change, ed. by B.E. Paulre, North-Holland, 1981.

13. Saunders, P.T., An Introduction To Catastrophe Theory, Cambridge University Press, Cambridge, 1980.

14. Thom, R., Structural Stability and Morphogenesis, English translation by D.H. Fowler of Stabilittie Structurelle et Morphogenese, Benjamin Reading, MA, 1975.

15. Zeeman, E.C., "Catastrophe Theory", Scientific American, vol. 234, April 1976.

16. Zeeman, E.C., Catastrophe Theory, Addison-Wesley, Reading, MA, 1976.

VITA

Jiefeng Qin was born on November 23, 1965 in Shanghai, People's Republic of China. He obtained his Bachelor's degree in Naval Architecture And Ocean Engineering in July 1988 from Shanghai Jiao Tong University, Shanghai, P.R. China. After graduation, he worked 2 years as an assistant engineer for Shanghai Merchant Ship Design And Research Institute. Graduate studies at Virginia Polytechnic Institute And State University began in August 1990, and a Master of Science degree in Civil Engineering was completed in April 1992. His career interests are in traffic engineering and transportation planning. He is a member of the Honor Society of Phi Kappa Phi.