Rural Networks:

Village Categorization -

An Important Basis For Forecasting

Mr. H. Leijon, ITU

UNION INTERNATIONALE DES TELECOMMUNICATIONS
INTERNATIONAL TELECOMMUNICATION UNION
UNION INTERNACIONAL DE TELECOMUNICACIONES

RURAL NETWORKS:

VILLAGE CATEGORIZATION - AN IMPORTANT BASIS FOR FORECASTING

by

Herbert Leijon

Legend

SLEPT	=	Village category variable, composed of sub-variables S,L,E,P,T	
S	=	Size	S=0: - 100 (population) 1: 100 - 500 2: 500 - 3000 3: 3000 - 10 000 4: 10 000 -
L	=	Level	L=0: (Almost) no functions 1: Basic, e.g., police,fire,nurse,elementary school... 2: Many functions for own day-to-day needs 3: Administrative and commercial centre for other towns (e.g., for the whole community) 4: Administrative and commercial centre for large area (e.g., for the whole district)
E	=	Socio-Economic type	E=0: Basic, e.g., Agriculture, fishing,.... 1: Mixed, e.g., Basic (i.e., as E=0 above) + small industries and businesses 2: Mixed, (i.e., as E=1 above) + large industries, businesses, administration
P	=	Private economic level	P=0: Poor 1: Average 2: High
T	=	Population development trend	T=0: Rapidly decreasing 1: Slowly decreasing 2: Constant 3: Slowly increasing 4: Rapidly increasing
POP	=	Population	
D	=	Density, corresponding to total demand	
DMAX	=	Density saturation limit for villages, corresponding to total demand	
C	=	Future number of connected subscriber lines	
PC	=	Proportion satisfied demand (connected lines / total demand)	
TCR	=	Total Calling Rate (traffic per subscriber line)	
PO	=	Proportion Originating traffic, out of total	
PI	=	Proportion Internal traffic, out of total	
v	=	Village	
c	=	commune	
d	=	district	
n	=	country	
t	=	point of time (0 = present year)	

Note: "Density" should correspond to "Total Demand" whether satisfied or not. A properly maintained waiting list may thus be included. This goes for past, present and future points of time. For a certain village **v** this means:

Number of connected main lines $C_v^{(t)} = POP_v^{(t)} \cdot D_v^{(t)} \cdot PC_v^{(t)}$

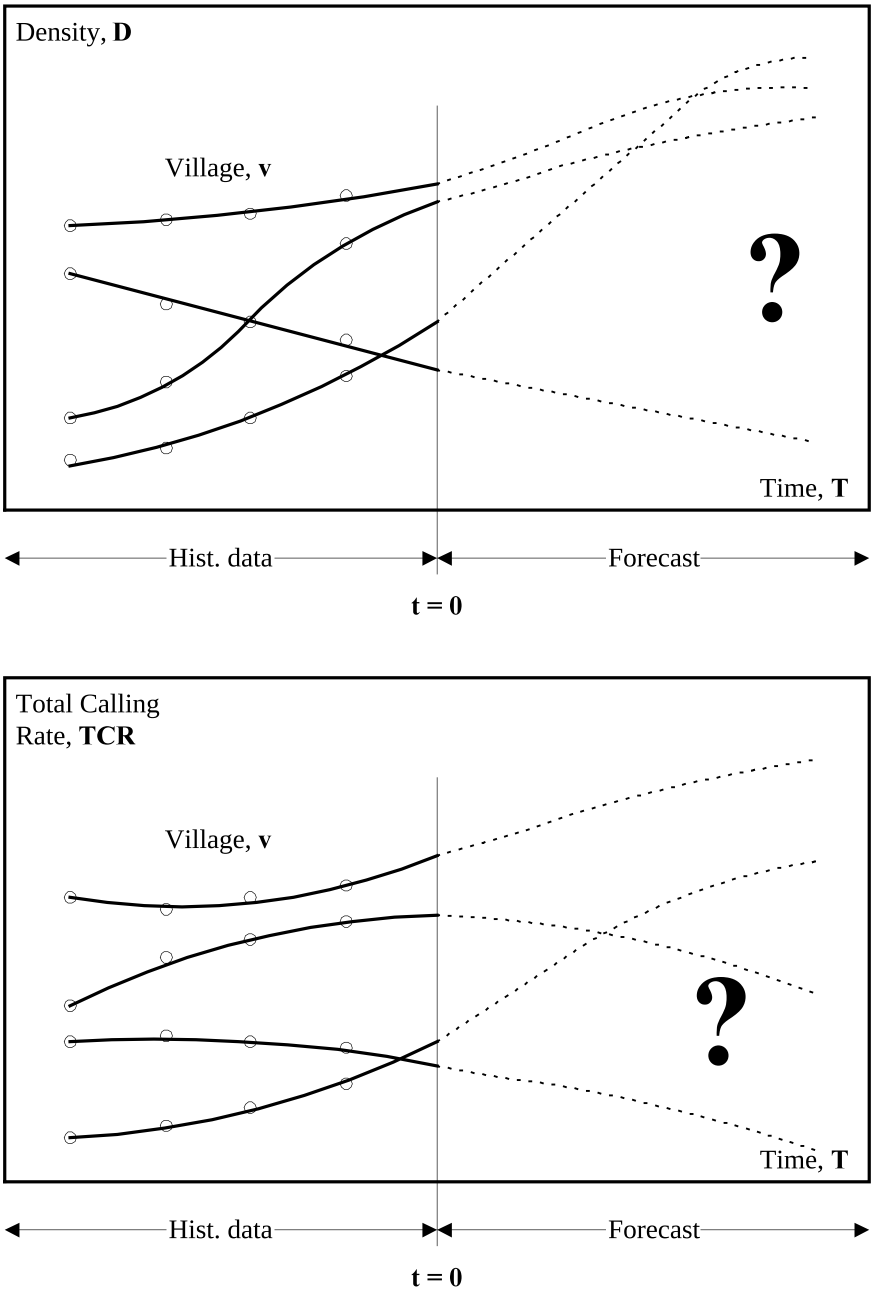
Total originating traffic = $C_v^{(t)} \cdot TCR_v^{(t)}$

For each rural area (e.g., called District) which is to be planned, we need to forecast:

a) the traffic interests between all sub-areas (e.g., called Communes);

b) the long distance traffics from and to each commune;

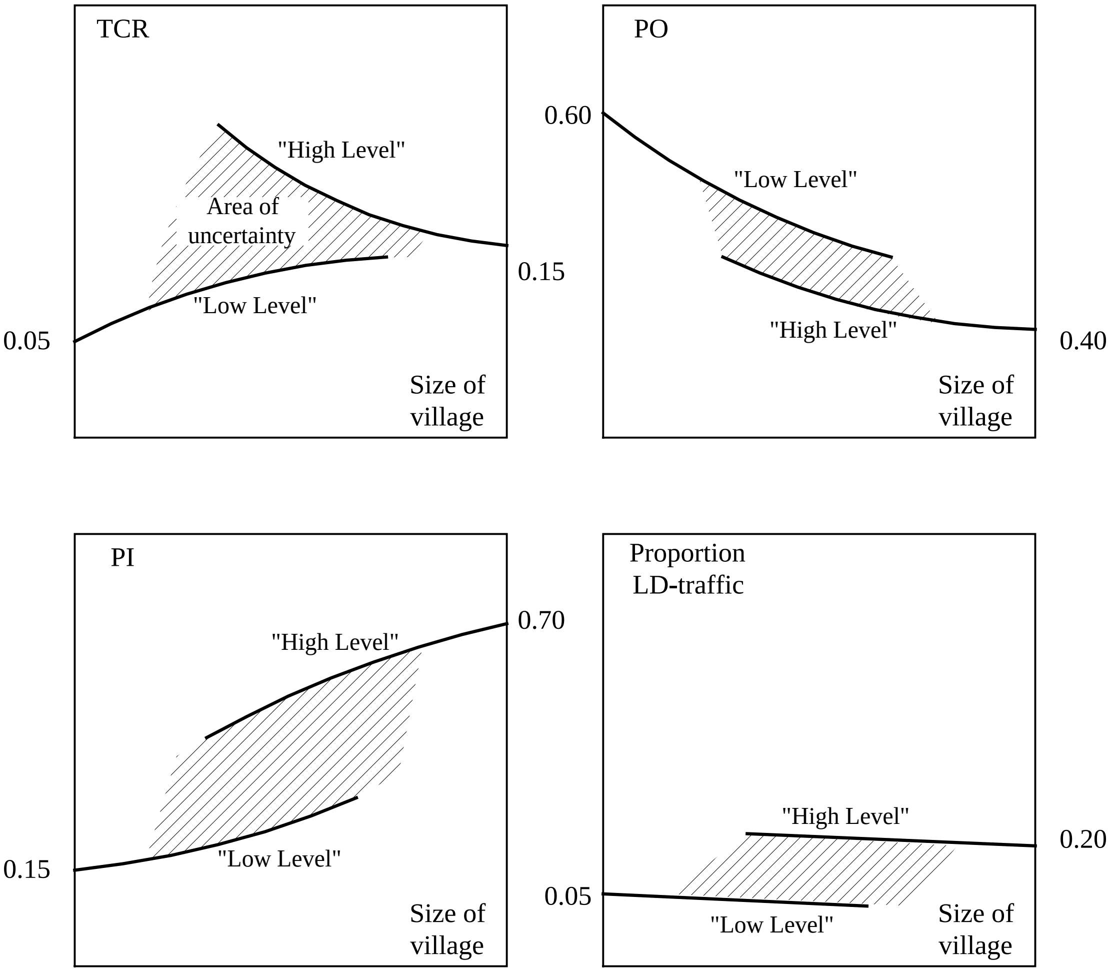
c) the number of subscribers (main lines) and the originating, terminating and internal traffic for each village and town in the rural area. For purposes of this paper, which concentrates on item c), "villages" covers both types.



Forecast problems (Forecast of PO and PI not shown)

Typically, there are about 10 to 20 communes in such a district, while there may be several hundred villages, and country-wide, there may be thousands of villages, but typically numbering about 10,000 to 50,000.

Furthermore, there are great differences between villages as regards population size, administrative and cultural level, socio-economic type, private-economic level, and population development trend.

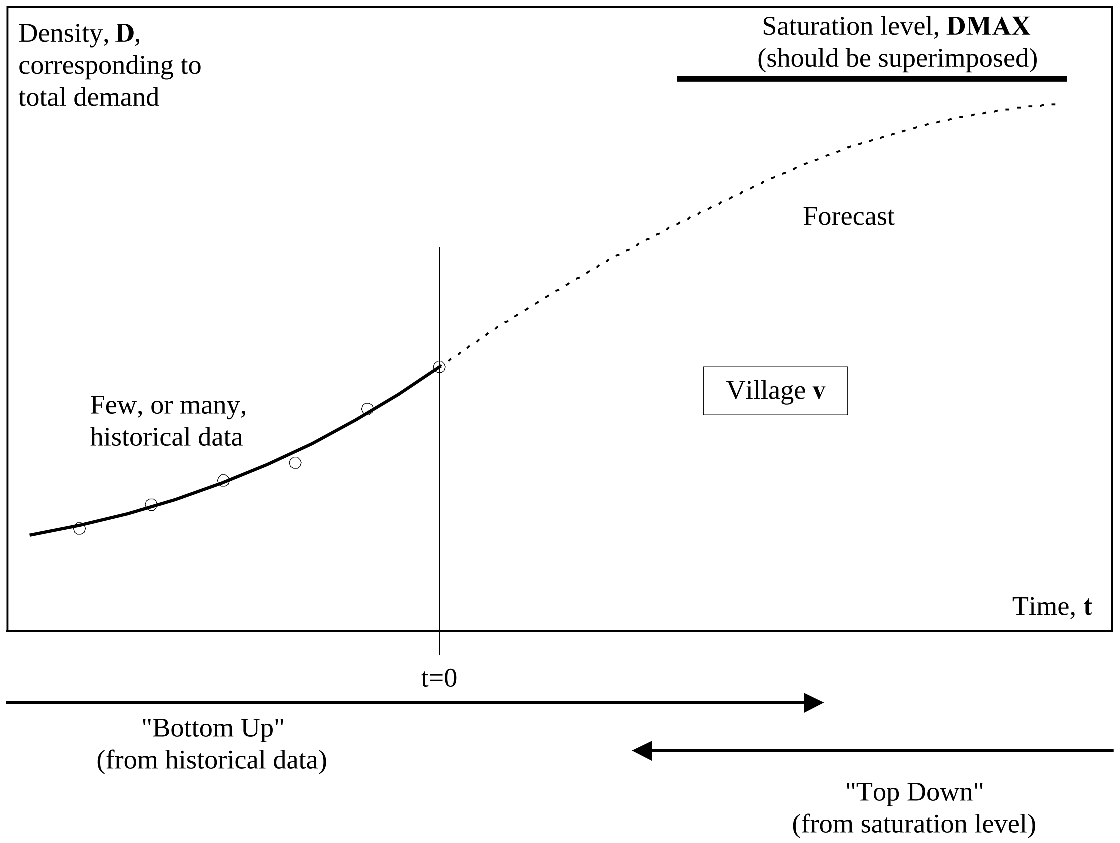


Examples of how traffic characteristics for villages in a particular area could be described, using only two parameters - Size of a village, and, "Level" of a village. "Level" would then be the substitute for several parameters (compare with L,E,P and T).
As a result of this rather rough way of describing the villages, areas of uncertainty would be quite large.

Consequently, we have a dilemma here; because of the differences between the villages, we ought to pay particular attention to village forecasting, yet because of the large number of villages, we need both to standardize and to computerize the village forecast procedure as much as possible, applying some kind of mathematical model.

One class of models that can be used to forecast the development of the main line density for such a diversified mass of entities is Growth Curves, e.g., the Exponential Logistic Model. Growth Curve models have a number of useful properties. One such property is that they function properly both if the supply of historical data is meagre or, on the contrary, if there are extensive data sets. Another welcome feature, is that when well used, Growth Curves can be used in a dynamic Bottom-Up/Top-Down process provided that a separately estimated saturation limit is superimposed on each curve, The basic growth curve function is namely in itself trend extrapolating forwards on the time axis (=Bottom-Up), while the superimposed saturation limit works backwards (=Top-Down).

The saturation limit should NOT, consequently, be a RESULT of the application of the growth curve model - that would change the process from this stable and reliable Bottom-Up/Top-Down method into a rather poor, uncontrollable, and unstable method.



How to forecast using Growth Curves:
Step 1) Forecast Saturation Level DMAX, separately;
Step 2) Fit a curve (trend line) to Historical Data <u>and</u> to DMAX.

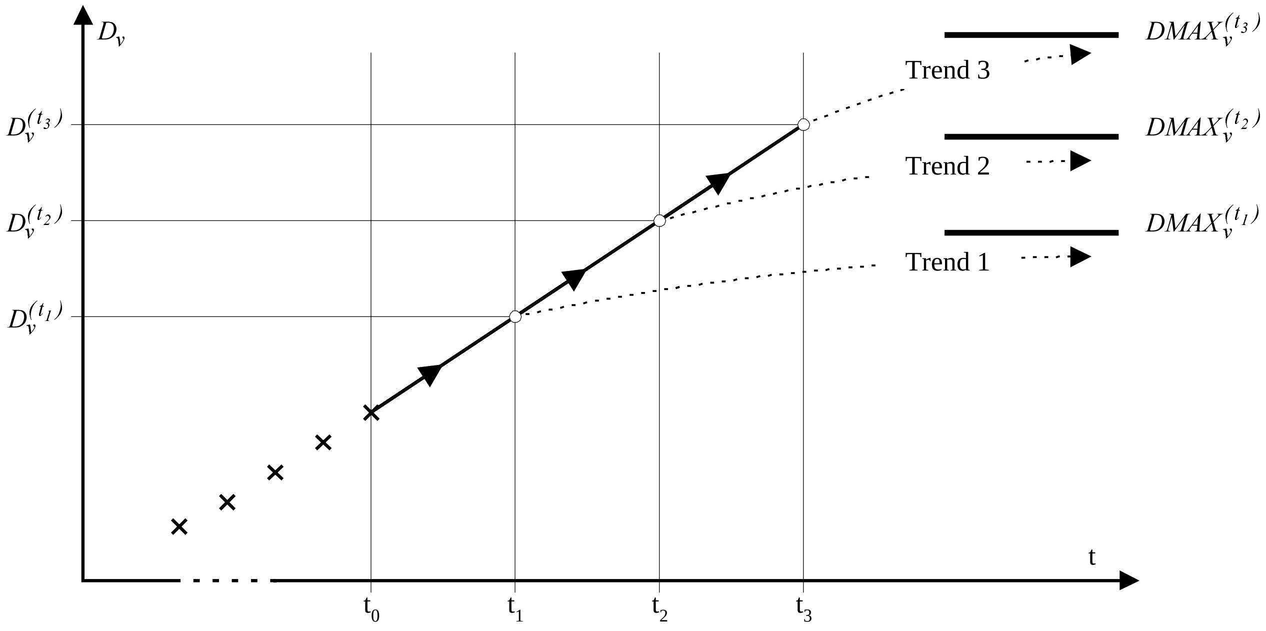
We can usually predict the development of the population or the number of households quite well. What we therefore need besides calling rate forecasts, is a forecast of density or penetration for each village.

Since we assume that there are at least some historical data available on densitites per village, $\boldsymbol{D}_{v}^{(t\leq 0)}$, we will need only the future density saturation limit for the same villages, $\mathbf{DMAX}_{v}$, to estimate the future densitites, $\boldsymbol{D}_{v}^{(t>0)}$, which will merely be a matter of straight forward numerical calculation according to the growth curve model.

What about **DMAX**, then?

Well, since the saturation limit **DMAX** mostly (although not absolutely necessarily) is meant as an <u>asymptotic</u> limit for growth curves, thus being defined for unlimited time, it should then in principle remain constant for all points of time **t**, i.e. **t** would not be a parameter. "$\mathbf{DMAX}_{v}$" would thus be the appropriate notation.

However, since any particular village may change character over time, leading to changed development trend, we need to modify **DMAX** also. Therefore, the variable is defined as $\boldsymbol{DMAX}_{v}^{(t)}$, thus being an instrument not only to calculate one particular trend curve for each different village, but also to control trend changes.



As an example, this figure aims to illustrate how the forecast for a particular village might be done, utilizing three different trend lines, to be used e.g. in case of two trend shifts.

The way of modifying $\boldsymbol{DMAX_v^{(t)}}$ for different points of time **t** is to change $\boldsymbol{SLEPT_v^{(t)}}$ and all associated parameters.

In a situation where trend changes for the whole area **x** can be foreseen, the average saturation level $\boldsymbol{DMAX_x^{(t)}}$ may be used as the instrument to modify all village forecasts simultaneously (and collectively). These two instruments, $\boldsymbol{DMAX_v^{(t)}}$ and $\boldsymbol{DMAX_x^{(t)}}$, may be used in combination (producing a sort of development trend for the whole area but allowing deviations for a few individual villages).

The average future saturation limits $\boldsymbol{DMAX_x^{(t)}}$ for a whole area **x** should be defined and used as an input to the algorithms that are aimed to estimate the village saturation limits $\boldsymbol{DMAX_v^{(t)}}$. Preferably, the variable should be defined for the whole rural area, i.e. the district (**x** = **d**). If defined for each commune (**x** = **c**) or for the whole country (**x** = **n**), the algorithms have to be modified.

As already mentioned above, our final goal is to forecast the future number of connected lines and the originating, terminating, and internal traffic, for each village.

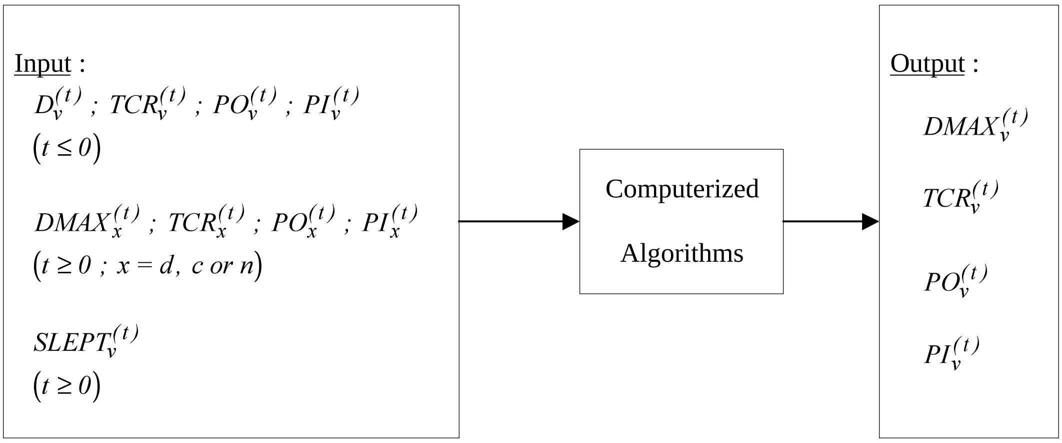
Once future population development is estimated, and since traffic quantities are calculated as the product of calling rates and connected main lines, we need only to forecast:

- future village density saturation limit $\boldsymbol{DMAX_v^{(t)}}$;
- future calling rate $\boldsymbol{TCR_v^{(t)}}$;
- future proportion of originating traffic $\boldsymbol{PO_v^{(t)}}$;
- future proportion of internal traffic $\boldsymbol{PI_v^{(t)}}$.

How can we do this? Well, first of all, we should of course use all relevant historical data about densities and calling rates. But in addition, we should also use our knowledge and well-grounded beliefs about character status and future of each village!

If we are able to describe each village in a district well enough we can then digitalize the descriptions by transferring them into the form of a numerical village category variable, $\boldsymbol{SLEPT_v^{(t)}}$.

Computer algorithms to estimate the required forecast variables from given village information can then be constructed.


Input :
$D_v^{(t)}$; $TCR_v^{(t)}$; $PO_v^{(t)}$; $PI_v^{(t)}$
$(t \leq 0)$
$DMAX_x^{(t)}$; $TCR_x^{(t)}$; $PO_x^{(t)}$; $PI_x^{(t)}$
$(t \geq 0\ ;\ x = d,\ c\ or\ n)$
$SLEPT_v^{(t)}$
$(t \geq 0)$
Computerized
Algorithms
Output :
$DMAX_v^{(t)}$
$TCR_v^{(t)}$
$PO_v^{(t)}$
$PI_v^{(t)}$

Estimation of forecast variables

So we have now two immediate tasks:

- to analyse the logical relationships between various significant village characteristics and required output from the process illustrated above;

- to construct the algorithms which are to operate on $\boldsymbol{SLEPT_v^{(t)}}$ and on $\boldsymbol{DMAX_x^{(t)}}$, and, of course, to define all other variables that are to be used in these algorithms.

We can easily understand that behind the demand for telecommunication services, there must be some driving forces. Depending on the means available, this demand might be satisfied by the installing of new main lines. Whether the new subcribers will use the newly provided services extensively or not, may also be a matter of the available means to satisfy the demand (one example of such means could be household economy). So we should examine the driving forces that could create a demand and the means that can help to satisfy this demand. Now reality, of course, is very complicated. Not only are the factors we need to look at interdependent rather than independent, but they are extremely numerous, and many are, in practice, next to impossible to quantify.

But if we can make a number of generalizations about a village, such as: the village is "relatively large"; its administrative, commmercial and cultural level is "high"; the socio-economic level is "pretty advanced"; the people living there are, in general, of a "rather good" economic level, and that the village seems to have a "good chance" to grow larger "rapidly", then we have a number of assumptions which we can plug into a model which will generate probable future scenarios upon which we can base decisions about telecommunication infrastructure planning.

If we now were to base our judgement about the future telecommunication development in the village upon these verbal generalizations, we would probably say that we expect a "fairly large" density growth leading to "a high" future density and that the usage of the provided services will also grow up to "a relatively large figure".

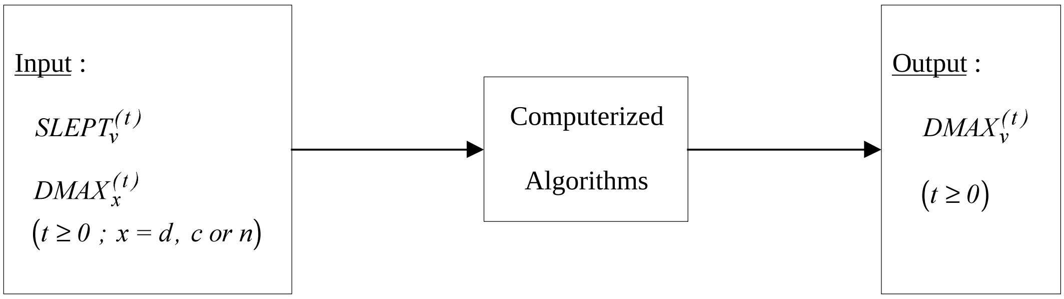
What we probably mean then is that this village is a quite dynamic one, filled of social and business activities, even to some extent dominating the surrounding area; in other words, we believe that both the driving forces behind the telecommunications demand and the means to satisfy the demand are quite substantial.

Maybe we are then ready to apply our ideas using the approach of village categorization. So we define numerical values for all parameters $\boldsymbol{S_v^{(t)}}$, $\boldsymbol{L_v^{(t)}}$, $\boldsymbol{E_v^{(t)}}$, $\boldsymbol{P_v^{(t)}}$ and $\boldsymbol{T_v^{(t)}}$ for each village **v** and for each point of time $\boldsymbol{t \geq 0}$. The settings for **t = 0** will be used to check the model parameters, thus giving us a good basis for the forecasting. The overall density saturation limit $\boldsymbol{DMAX^{(t)}}$ should also be defined, probably for the whole district, i.e. $\boldsymbol{DMAX_d^{(t)}}$.

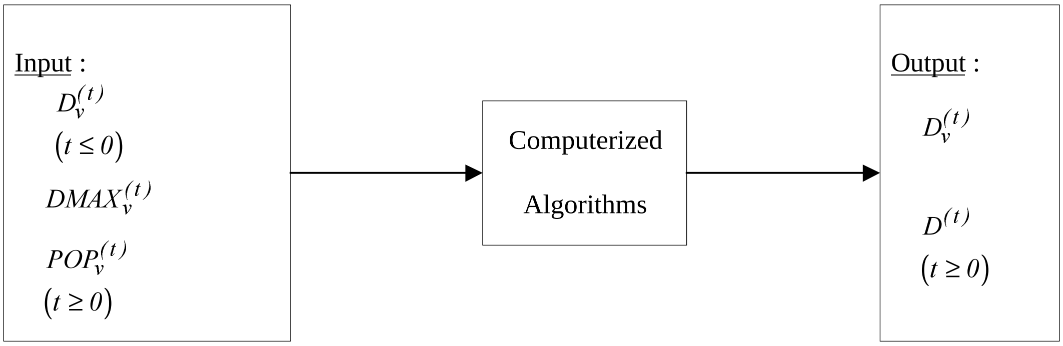
If we for the moment neglect necessary checks and corresponding adjustments and iterations, our little flowchart to determine $\boldsymbol{DMAX_v^{(t)}}$, $\boldsymbol{TCR_v^{(t)}}$, $\boldsymbol{PO_v^{(t)}}$ and $\boldsymbol{PI_v^{(t)}}$ may now be figured out more in detail and may be divided into three consequtive sub-flowcharts as per below:

Sub-Flowcharts:

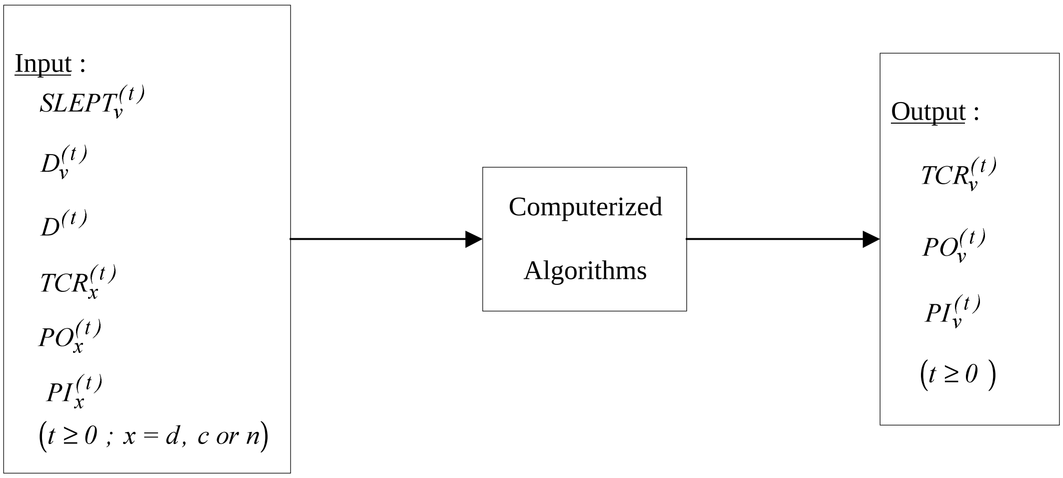
A



B



C



Sub-flowchart A:
<u>To calculate saturation levels per village</u>

Say that $\boldsymbol{DMAX_x^{(t)}}$ is the average saturation level in the area, e.g. the district (**x = d**).

If a particular village **v** now is an absolutely "average" one from all significant aspects, then we should expect that the saturation level to be used for this village should be the same as the average level, i.e. that $\boldsymbol{DMAX_v^{(t)} = DMAX_x^{(t)}}$.

If on the other hand the village differs significantly from the "average" village in a way reflected by the parameters $\boldsymbol{SLEPT_v^{(t)}}$, we want to calculate the specific saturation level $\boldsymbol{DMAX_v^{(t)}}$ using factors based on $\boldsymbol{SLEPT_v^{(t)}}$.

A simple formula for such an estimation may be:

$$DMAX_v^{(t)} = DMAX_x^{(t)} \cdot FDS_v^{(t)} \cdot FDL_v^{(t)} \cdot FDE_v^{(t)} \cdot FDP_v^{(t)} \cdot FDT_v^{(t)}$$

where e.g.

FDS = **F**actor for estimation of max **D**ensity based on **S**ize of the village;

FDL = **F**actor for estimation of max **D**ensity based on **L**evel of the village;

etc

In the case of an "average" village, all factors ***FDS***, ***FDL***,... should be = **1**, otherwise not.

Sub-flowchart B:
To calculate future densities per village

This is where we run our main Growth Curve algorithm, e.g. the Exponential Logistic model. We need two sets of data:

a) Past and present densities per village = Historical data;
b) Future saturation levels per village = Output of sub-flowchart **A**.

The calculation is a matter of fitting a curve to historical data under the constraint that the **imposed** saturation level must be the asymptote to the extrapolation of the curve.

From the resulting future densities per village $D_v^{(t)}$ we maycalculate the mean future densities per commune or for the whole district.

Sub-flowchart C:
To calculate future calling rates per village

$TCR_x^{(t)}$, $PO_x^{(t)}$ and $PI_x^{(t)}$ are the average calling rate, the average proportion of originating traffic, and the average proportion of internal traffic in the area, e.g. the district (**x** = **d**).

Again, if the particular village **v** is absolutely "average", we should then expect that these average values for the whole area would hold good also for the village considered; if not, we want to calculate the specific calling rates $TCR_v^{(t)}$, $PO_v^{(t)}$ and $PI_v^{(t)}$ using factors based on $SLEPT_v^{(t)}$.

A set of simple formulas to determine the individual village parameters could be as follows:

$$TCR_v^{(t)} = TCR_x^{(t)} \cdot FTS_v^{(t)} \cdot FTL_v^{(t)} \cdot FTE_v^{(t)} \cdot FTP_v^{(t)} \cdot FTT_v^{(t)} \cdot FTD_v^{(t)};$$

$$PO_v^{(t)} = PO_x^{(t)} \cdot FOS_v^{(t)} \cdot FOL_v^{(t)} \cdot FOE_v^{(t)} \cdot FOP_v^{(t)} \cdot FOT_v^{(t)} \cdot FOD_v^{(t)};$$

$$PI_v^{(t)} = PI_x^{(t)} \cdot FIS_v^{(t)} \cdot FIL_v^{(t)} \cdot FIE_v^{(t)} \cdot FIP_v^{(t)} \cdot FIT_v^{(t)} \cdot FID_v^{(t)}.$$

where e.g.

FTS = Factor for estimation of **t**otal calling rate based on **S**ize of the village;

FOS = Factor for estimation of proportion **O**riginating traffic based on **S**ize of the village;

FIS = Factor for estimation of proportion **I**nternal traffic based on **S**ize of the village;

FTD = Factor for estimation of **t**otal calling rate based on **D**ensity in the village;

etc.

In the case of an "average" village, all factors **FTS**, **FTL** ... **FOS**, **FOL** ... **FIS**, **FIL** ... should be = **1**, otherwise not.

Note that $\boldsymbol{TCR_x^{(t)}}$, $\boldsymbol{PO_x^{(t)}}$ and $\boldsymbol{PI_x^{(t)}}$ are input values to the algorithms. However, after having run the algorithms, thus obtaining $\boldsymbol{TCR_v^{(t)}}$, $\boldsymbol{PO_v^{(t)}}$ and $\boldsymbol{PI_v^{(t)}}$ as an output, we can calculate new values $\boldsymbol{TCR_x^{(t)}}$, $\boldsymbol{PO_x^{(t)}}$ and $\boldsymbol{PI_x^{(t)}}$ from this output. Therefore, the whole calculation may be iterative!

From SLEPT to FDS

We have seen that the calculation of future densities and future calling rates involves the use of a number of "factors" $\boldsymbol{FDS_v^{(t)}}$, $\boldsymbol{FDL_v^{(t)}}$,...etc, which are based on $\boldsymbol{SLEPT_v^{(t)}}$.

Having first defined numerical values of $\boldsymbol{SLEPT_v^{(t)}}$ $(\boldsymbol{t \geq 0})$, our problem will consequently be to translate these data into "factors".

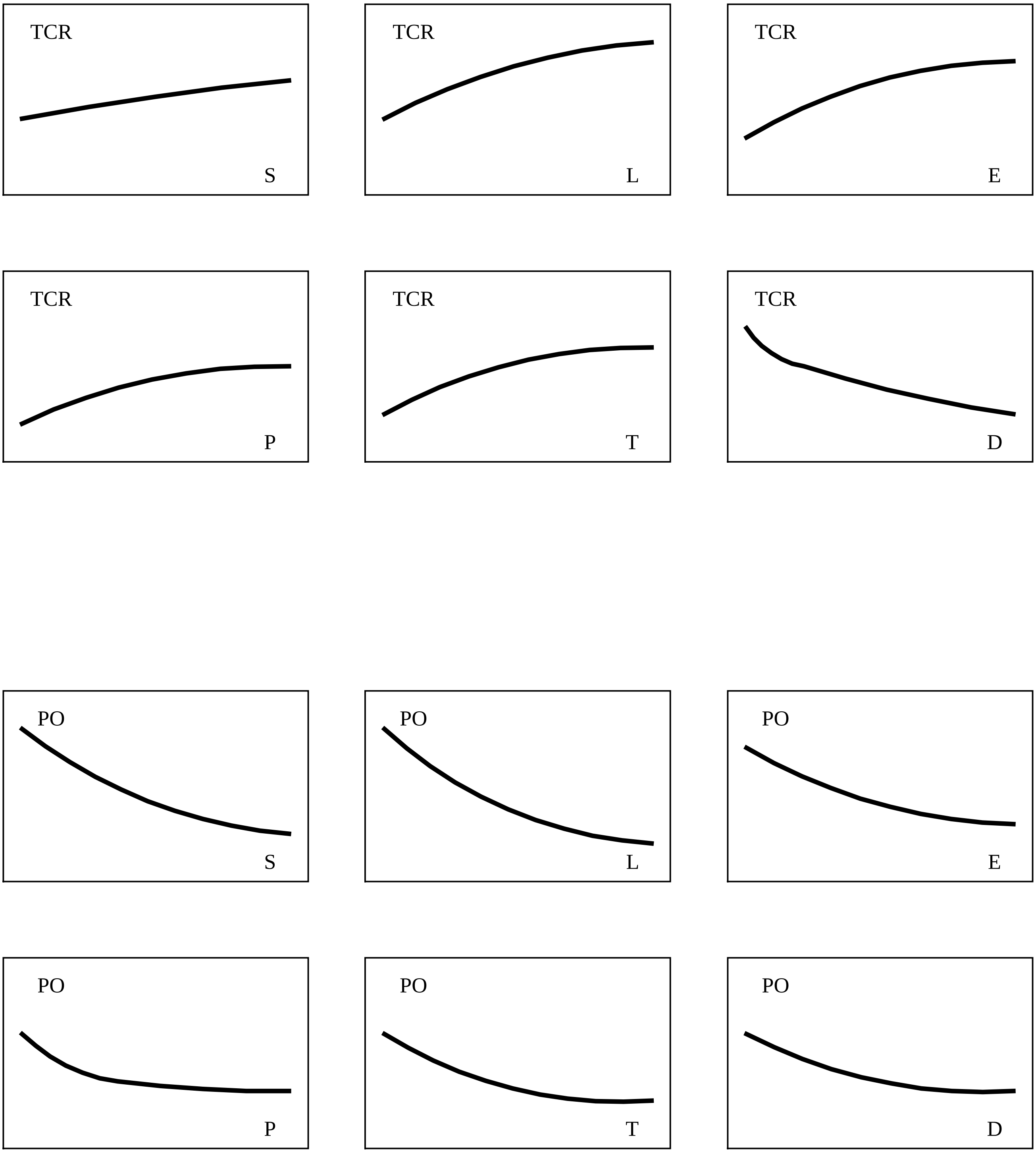
Let us then first study how the needed parameters **TCR**, **PO**, **PI** and **DMAX** are influenced by **S**, **L**, **E**, **P**, **t** and **D**.
Take **TCR**, for example. Maybe this parameter is most strongly affected by the **L**evel of the village, i.e. parameter **L**. Maybe the next strongest influence comes from the socio-**E**conomic type, i.e. parameter **E**, etc.

If we represent this by figures, so that "**1**" means the strongest influence, "**2**" means the next to strongest influence, etc, then we might demonstrate our ideas in form of a table as per below:

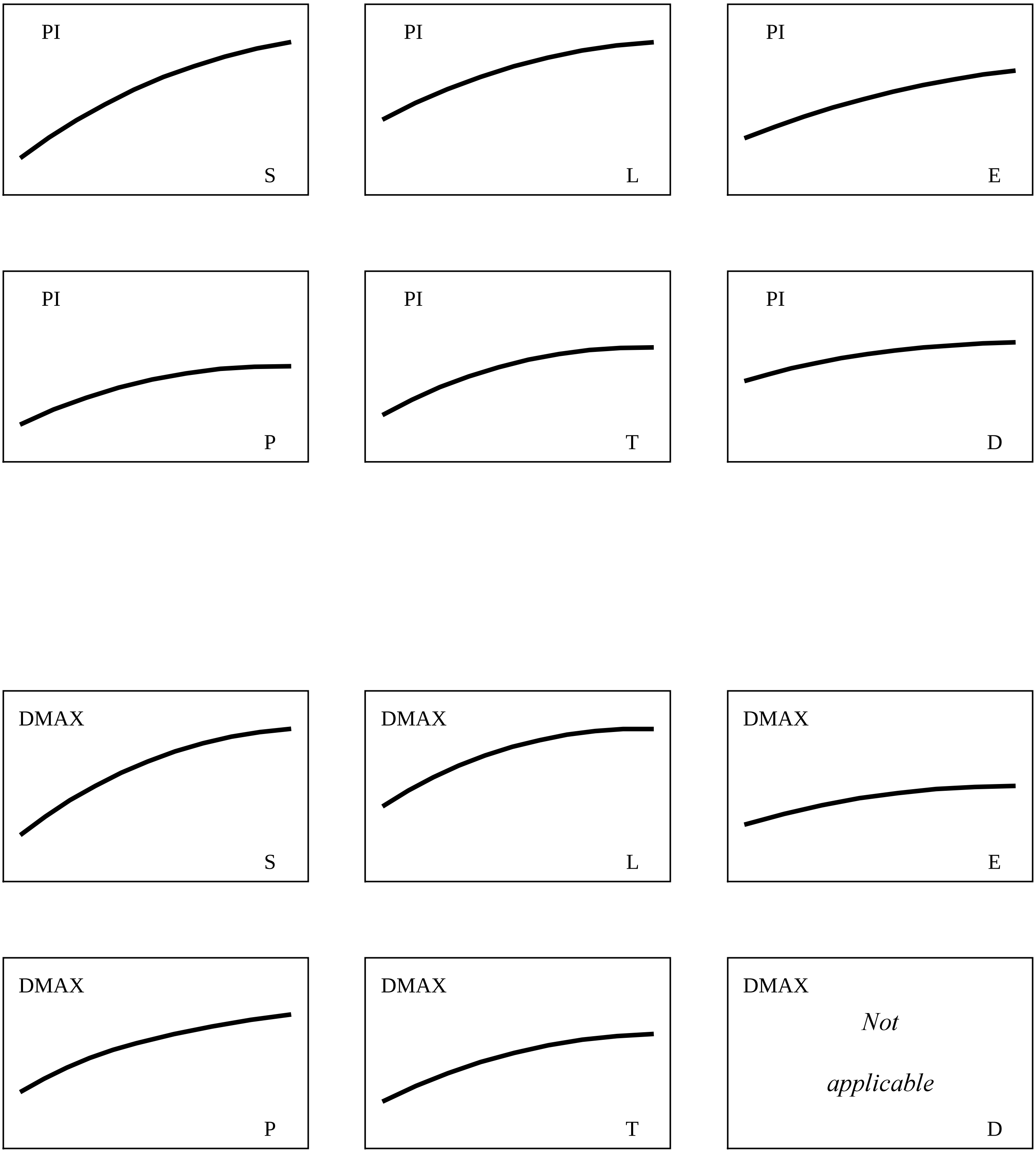
Influence on by	**TCR**	**PO**	**PI**	**DMAX**
S	5	2	1	1
L	1	1	2	2
E	2	3	3	3
P	3	5	4	4
T	6	4	5	5
D	4	6	6	Not applicable

*General influence on **TCR**, **PO**, **PI** and **DMAX** by **S**, **L**, **E**, **P**, **t** and **D**.*
"1" represents strongest influence, "2" next-to strongest, etc.
Note: Given figures are just examples!

After this exercise, we may be ready to express our ideas on "influence" graphically, but still in a qualitative manner, i.e. not introducing quantities (actual numerical values) as yet.



General influence on ***TCR*** *and* ***PO*** *by* ***S****,* ***L****,* ***E****,* ***P****,* ***T*** *and* ***D****, expressed graphically.*



General influence on ***PI*** *and* ***DMAX*** *by* ***S****,* ***L****,* ***E****,* ***P****,* ***T*** *and* ***D****, expressed graphically.*

Inspecting all these curves, we find that they generally may be characterizedby two parameters:

B = Inclination of the curve;
Z = Shape of the curve

Concerning the **inclination** of a curve **B**, we see that a curve may be nearly horizontal (weak influence), positively steeper (larger, positive influence), or negatively steeper (larger, negative influence).

So we could use **B** to characterize the inclination as per below:

{B} =

Influence **on** **by**	**TCR**	**PO**	**PI**	**DMAX**
S	+ 3	- 1	+ 1	+ 1
L	+ 1	- 2	+ 2	+ 2
E	+ 1	- 2	+ 3	+ 3
P	+ 2	- 3	+ 3	+ 3
T	+ 3	- 3	+ 3	+ 3
D	- 2	- 3	+ 3	Not applicable

B = **Steepness** of curve, indicating degree of influence;

|B| = **1** represents **largest** inclination;

|B| = **2** represents **next-to largest** inclination; etc.

(**|B|** = **absolute value** of **B**)

+ represents **positive** inclination;

- represents **negative** inclination.

Note: Given figures are just <u>examples</u>!

Let us now look at the other property - the shape of a curve, **Z**.

Wee see that all curves are more or less bent, some of them being **convex**, others **concave**, and bent to different **degrees**, from slightly bent to quite strongly bent. Each particular curve has not a constant bend, but the bend decreases with increasing value of the influencing parameter.

We see also that all curves that have positive inclination are bent alike, say concavely, and that all curves having negative inclination are bent the other way - convexly.

We may use **Z** to characterize the shape of curves as per below:

{Z} =

Influence **on** **by**	**TCR**	**PO**	**PI**	**DMAX**
S	2	3	2	2
L	3	2	2	2
E	2	1	2	1
P	1	1	1	1
T	1	1	1	1
D	2	2	1	Not applicable

Z = Shape of curve;
Z = **1** represents less bent curve;
Z = **2** represents medium bent curve;
Z = **3** represents more bent curve.

Note: Given figures are just **examples**!

Any curve may then be calculated as

$$FYQ = \frac{|B| \cdot CB + \left(\sqrt[Z+CZ]{Q} - \sqrt[Z+CZ]{\overline{Q}} \right) \cdot TB}{|B| \cdot CB}$$

where

Y = letter **D**, **T**, **O** or **I** (representing **DMAX**, **TCR**, **PO** or **PI**);

Q = **S**, **L**, **E**, **P**, **T** or **D**;
(**Q** is interpreted as a letter in the name "**FYQ**", otherwise as a variable)

$\overline{\mathbf{Q}}$ = average value of **S**, **L**, **E** ,**P**, **T** or **D**;

B = curve steepness parameter;

Z = curve shape parameter;

CB = scale constant for **B**;

CZ = scale constant for **Z**;

TB = **+1** for positive **B**-values, **-1** for negative **B**-values;

|B| = absolute value of **B**.

Interpretation of $\overline{\mathbf{Q}}$:

$$\overline{S} = 2\ ;\ \overline{L} = 2\ ;\ \overline{E} = 1\ ;\ \overline{P} = 1\ ;\ \overline{T} = 2$$

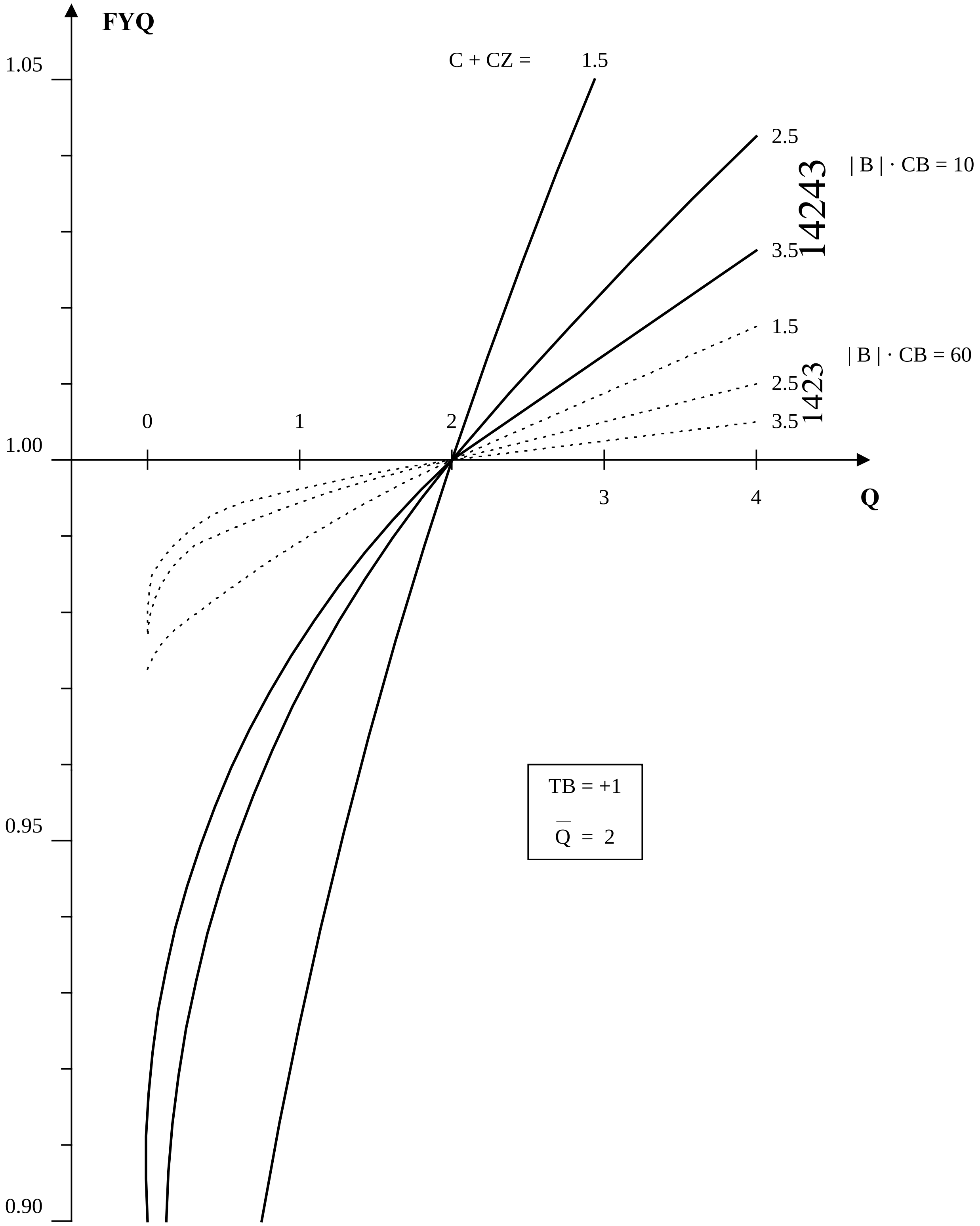
$$\overline{D} = \frac{\sum_{v} POP_v \cdot D_v}{\sum_{v} POP_v}$$

Interpretation of Scale Constants:

CB and **CZ** may be used to manipulate a whole range of curves simultaneously.
Order of magnitude:

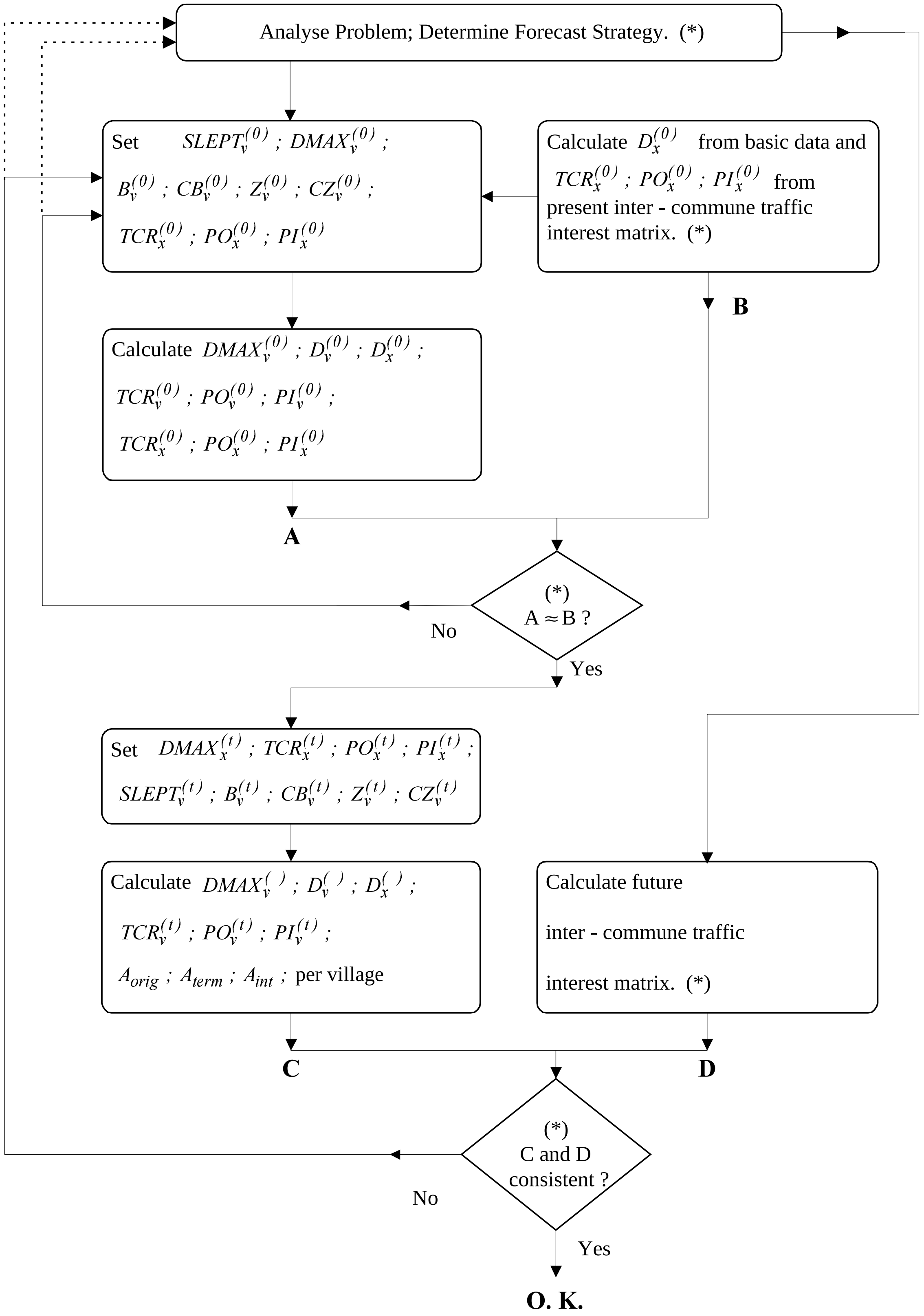
$$CB \approx 20;\ \ CZ \approx 0.5$$

Note: A smaller value of $|B| \cdot CB$ indicates a steeper curve;
A smaller value of $Z+CZ$ indicates a less bent curve. ($Z+CZ = 1$ will produce a straight line)



Demonstration of the effect on ***FYQ*** *of the two quantities* $|\boldsymbol{B}|\cdot$ ***CB*** *and* ***Z+CZ****, for different* ***Q****-values.*

Now we are ready to run the algorithms. We should then follow some kind of logical flowchart, like the one shown below.


Analyse Problem; Determine Forecast Strategy. (*)
Set $SLEPT_v^{(0)}$; $DMAX_v^{(0)}$; $B_v^{(0)}$; $CB_v^{(0)}$; $Z_v^{(0)}$; $CZ_v^{(0)}$; $TCR_x^{(0)}$; $PO_x^{(0)}$; $PI_x^{(0)}$
Calculate $D_x^{(0)}$ from basic data and $TCR_x^{(0)}$; $PO_x^{(0)}$; $PI_x^{(0)}$ from present inter - commune traffic interest matrix. (*)
Calculate $DMAX_v^{(0)}$; $D_v^{(0)}$; $D_x^{(0)}$; $TCR_v^{(0)}$; $PO_v^{(0)}$; $PI_v^{(0)}$; $TCR_x^{(0)}$; $PO_x^{(0)}$; $PI_x^{(0)}$
B
A
(*) A ≈ B ?
No
Yes
Set $DMAX_x^{(t)}$; $TCR_x^{(t)}$; $PO_x^{(t)}$; $PI_x^{(t)}$; $SLEPT_v^{(t)}$; $B_v^{(t)}$; $CB_v^{(t)}$; $Z_v^{(t)}$; $CZ_v^{(t)}$
Calculate $DMAX_v^{(\ )}$; $D_v^{(\ )}$; $D_x^{(\ )}$; $TCR_v^{(t)}$; $PO_v^{(t)}$; $PI_v^{(t)}$; A_{orig} ; A_{term} ; A_{int} ; per village
Calculate future inter - commune traffic interest matrix. (*)
C
D
(*) C and D consistent ?
No
Yes
O. K.

Forecast scheme for rural networks.
**) The main problem analysis, calculations on the inter-commune traffic matrix, and decisions and comparisons between village and inter-commune traffic data will be treated in another document.*

When you study this flowchart, you may be astonished to see that some parameters appear both as input and output, and that sometimes a comparison or check is done between them!

The explanation to that is as follows:

Basic data are generally aggregate data. It is not possible to disaggregate these data in a mathematically unambiguous way.

The calculations we do are contrary to that quite detailed (disaggregated) and are based on the combined use of basic, aggregate data, and detailed, hypothetical parameter values.

If we, after running the algorithms, aggregate our results, we may obtain sets of data that are compatible with the basic data. We can only hope that the numerical values are about the same. If not, we may have to reconsider the situation. Maybe the differences are quite acceptable, maybe not.