Dynamic Line Management

Problem presented by

Mike Stevens and Nick Kelly

Talk Talk Technology

Executive Summary

Talk Talk Technology has an ongoing programme of Dynamic Line Management (DLM), aimed at improving the performance of ADSL broadband connections. Choosing the most appropriate configuration, or 'profile', for an individual line allows a balance to be reached between achieving a high line speed and a stable customer experience. Among the objectives of DLM are to determine the likely line speed of new connections and to identify lines that might benefit from an updated profile.

Addressing these questions requires three complementary types of modelling: physical modelling of the electrical characteristics of ADSL lines (these are existing copper telephone lines); information modelling of the associated communication channel, taking into account the levels of noise and the coding of ADSL signals; and finally statistical modelling, which enables data gathered from the exchange to be used to construct relationships between key variables and to identify lines that might be faulty or on poor profiles. The work of the Study Group in April 2010 contributed mainly to developing the statistical modelling. This report summarises the progress that was made and the opportunities for developing these techniques further.

Version 1.1
June 14, 2010

Report authors

Robert Leese (Industrial Mathematics KTN)
Anirban Mondal (Texas A&M University)
Warren Smith (University of Birmingham)

Contributors

Andrew Lacey (Heriot-Watt University)
Anirban Mondal (Texas A&M University)
Charles Mberi Kimpolo (African Institute for Mathematical Sciences)
John Ockendon (University of Oxford)
Montaz Ali (Witwatersrand University)
Robert Leese (Industrial Mathematics KTN)
Robert Whittaker (University of Oxford)
Steven Hill (University of Warwick)
Siu Kwan Yip (University of Warwick)
Tomasz Brozek (Warsaw School of Information Technology)
Warren Smith (University of Birmingham)

ESGI73 was jointly organised by
The University of Warwick
The Knowledge Transfer Network for Industrial Mathematics

and was supported by
The Engineering and Physical Sciences Research Council

Contents

1 Introduction

1.1 Background

(1.1.1) Prior to taking a new customer, Talk Talk Technology provides an estimate on the likely line speed. This is an important part of the sale and the estimate needs to be as accurate as possible. Broadband speeds have been the subject of much consumer and Ofcom interest. One of the challenges with providing accurate speed estimates is that at the pre-sales stage only the line length, resistance and capacitance are made available. Currently the line length is used to determine the likely speed, but it is known that there are quality issues with the BT line length data.

(1.1.2) Talk Talk Technology posed three questions to the study group.

- Can a better estimate of speed against line length, resistance and capacitance be made?
- Is it possible to identify systematically the errors in supplied parameters (*e.g.* line length)?
- Can we distinguish between lines that are configured incorrectly (on the wrong profile) and faulty lines?

(1.1.3) The Study Group concentrated on two aspects of these problems. The first issue was to determine an improved estimate of line length at the pre-sales stage. The relationship between capacitance (from the wire to earth) and the line length was the focus of this study. The second issue was to investigate better estimates of the speed.

(1.1.4) The group's work involved three related strands: physical modelling, information modelling and the statistical modelling of the dataset provided by Talk Talk Technology. Section 2 below deals with physical and information modelling, Section 3 looks at relevant statistical modelling techniques, and Section 4 provides an initial application of these techniques to the Talk Talk dataset. Finally, Section 5 draws some conclusions and makes recommendations for further work.

2 Physical and information modelling

2.1 Capacitance

(2.1.1) The electrical properties of a twisted pair can be investigated analytically using a transmission line model, which is explored further in the next subsection. The Talk Talk dataset contains measurements of line capacitance, and so in this subsection we look at how capacitance may be calculated from first principles. Figure 1 shows a schematic cross-sectional representation of the wire geometry.

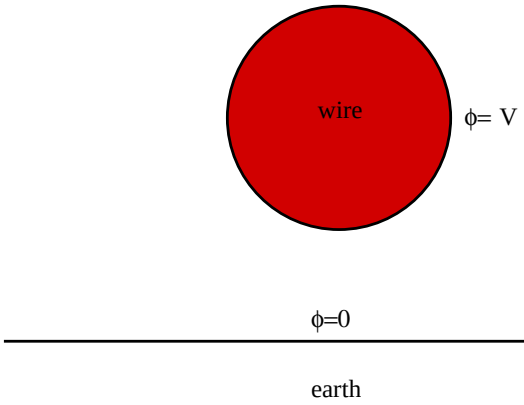

Figure 1: Schematic of the configuration of one wire and earth. The earth sits at zero electric potential, and the second wire is assumed to lie symmetrically below, with equal and opposite potential to the wire that is shown.

(2.1.2) The first stage in the evaluation of the capacitance per unit length is to determine the electric potential in the region between the wire and the earth. We define $\phi(x, y)$ to be the potential at an arbitrary point and V to be the constant potential on the surface of the wire. The governing equation for the potential is given by

$$\nabla^2 \phi = 0, \tag{1}$$

with boundary conditions

$$\phi = V$$

on the surface of the wire,

$$\phi = 0$$

on the earth and

$$|\nabla \phi| \to 0 \quad \text{as } |\boldsymbol{x}| \to \infty$$

in the far field.

(2.1.3) An elegant way forward is to transform this boundary-value problem using bipolar cylindrical coordinates

$$x = \frac{a \sinh(v)}{\cosh(v) - \cos(u)},$$

$$y = \frac{a \sin(u)}{\cosh(v) - \cos(u)},$$

for $u \in [0, 2\pi]$ and $v \in (0, v_1)$, where $v = 0$ is the boundary with the earth and $v = v_1$ is the surface of the wire. In this coordinate system, the

Laplacian becomes

$$\nabla^2\phi = \frac{\cosh(v)-\cos(u)}{a^2}\left(\frac{\partial^2\phi}{\partial u^2}+\frac{\partial^2\phi}{\partial v^2}\right)=0 \tag{2}$$

with the boundary conditions on $\phi(u,v)$ being

$$\phi(u,0)=0, \quad \phi(u,v_1)=V, \quad \frac{\partial\phi}{\partial u}(0,v)=0, \quad \frac{\partial\phi}{\partial u}(2\pi,v)=0.$$

(2.1.4) The solution is given by

$$\phi = \frac{Vv}{v_1}.$$

The surface charge is obtained by evaluating the normal derivative to the wire in the bipolar coordinate system

$$\sigma = \epsilon\frac{\partial\phi}{\partial n} = \frac{\epsilon}{h_v}\frac{\partial\phi}{\partial v} = \frac{\epsilon}{h_v}\frac{V}{v_1},$$

where h_v is the scale factor along the v axis and ϵ is the permittivity. The total charge on the wire is deduced by integrating the surface charge around the wire surface:

$$Q = \int_{\text{wire}}\sigma dS = \int_{u=0}^{2\pi}\epsilon\frac{\partial\phi}{\partial v}du = \frac{2\pi\epsilon V}{\cosh^{-1}(h/r)},$$

where h is the distance between the centre of the wire and the earth and r is the radius of the wire.

(2.1.5) We note that the scale factor from the surface charge and the integration cancel to produce the simplified result. Hence we obtain the capacitance per unit length (which we compare against measured data in upcoming statistical analysis) as

$$C_E = \frac{2\pi\epsilon}{\cosh^{-1}(h/r)}. \tag{3}$$

(2.1.6) This expression is well-established in the transmission line literature. See for example the short conference paper [5].

(2.1.7) The capacitance per unit length is expected to be constant. The dataset provided by Talk Talk Technology suggests that $C_E \approx 7\times10^{-11}\text{Fm}^{-1}$, which from (3) implies that $h/r \approx 1.3$.

2.2 Transmission line model

(2.2.1) A standard transmission line model for the line voltage $V(x,t)$ takes the form

$$\frac{\partial^2 V}{\partial x^2} = LC\frac{\partial^2 V}{\partial t^2} + RC\frac{\partial V}{\partial t}, \tag{4}$$

where L is the inductance per unit length, R the resistance of the wire per unit length, C the capacitance between the wires per unit length, x the distance along the line and where t is time.

(2.2.2) We wish to investigate the propagation of a mode with angular frequency ω, and so seek solutions to (4) of the form

$$V = e^{i\omega t} f(x),$$

which requires that $f(x)$ then satisfies the ordinary differential equation

$$f''(x) + (LC\omega^2 - RCi\omega)f(x) = 0. \tag{5}$$

(2.2.3) The solution of (5) is $f(x) = e^{i\lambda x}$, where

$$\lambda^2 = LC\omega^2 - RCi\omega.$$

Note that λ is complex-valued. The physically meaningful solution has $\mathrm{Im}\ \lambda > 0$, corresponding to a signal that attenuates as it propagates. A binomial expansion for large frequencies yields

$$\mathrm{Im}\ \lambda \ \sim \frac{1}{2}R\left(\frac{C}{L}\right)^{\frac{1}{2}} - \frac{1}{16}\left(\frac{R}{L\omega}\right)^3 (LC)^{\frac{1}{2}}\,\omega.$$

Hence we have an exponential decay of signal strength with distance, with, as expected, the highest frequencies suffering more attenuation.

(2.2.4) Typical values of the parameters in this model are $L = 0.6\mathrm{mH/km}$, $C = 38\mathrm{nF/km}$ and $R = 300\Omega/\mathrm{km}$ (although note that a more detailed model incorporates a frequency dependence in R and L [5]). These values give $LC \approx 2\times10^{-17}\mathrm{m^{-2}s^2}$ and $RC \approx 10^{-11}\mathrm{m^{-2}s}$. If we consider a frequency of 100kHz, corresponding to an angular frequency ω of approximately $6.3\times10^5\mathrm{s^{-1}}$, then we find $\lambda \approx (7-6i)^{1/2}10^{-3}\mathrm{m^{-1}}$ and $\mathrm{Im}\lambda \approx 1\times10^{-3}\mathrm{m^{-1}}$. Therefore a signal with frequency of 100kHz attenuates by a approximately one-third every kilometre.

(2.2.5) We can plot the factor by which modes of different frequencies are attenuated after transmission of one kilometre. Figure 2 shows this variation.

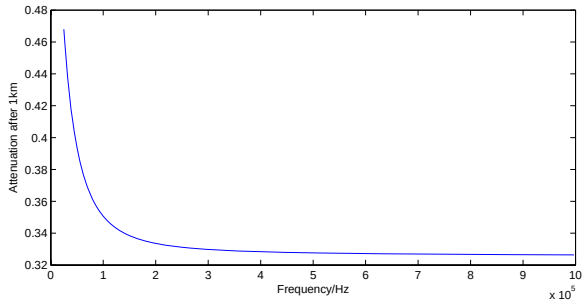

Figure 2: The attenuation of the signal as a function of the frequency after 1km.

2.3 Line speeds and information modelling

(2.3.1) Signal attenuation with increasing line length is a major factor in determining line speeds, but one must also consider the information-theoretic aspects. Each ADSL 'tone' acts as a narrowband subchannel of 4kHz, and the total data rate will be the sum of contributions from the individual tones. The data rate for a single tone (labelled by k) is approximately [5]

$$C_k = \frac{1}{T}\log_2(1 + \frac{\mathrm{SNR}_k}{\Gamma})\,, \tag{6}$$

where $1/T$ is the data symbol rate on each tone, SNR_k is the signal-to-noise ratio for the particular tone, and Γ is a factor that incorporates the noise margin and coding gain and the requirements for bit error rates.

(2.3.2) This expression assumes that the power spectral densities of the signal and the noise are flat within each tone, which is justified by the fact that the individual tones are narrow. However, they may vary appreciably across the full set of tones (ADSL has 256 tones and ADSL2+ has 512 tones). This variation is one source of the difference in signal-to-noise ratio between different tones. Another is the frequency-dependent signal attenuation that we have already seen.

(2.3.3) In order to estimate the total data rate (*i.e.* the line speed), by summing the data rates for each individual tone, one would need more information about the power spectral densities for signal and noise across the full frequency band, and also knowledge of the coding being used, so that the factor Γ can be set accordingly. An illustration of how the calculation works is given in [2].

(2.3.4) Even without more detailed analysis, it is clear that the fall-off of line speed with line length does not generally follow the exponential attenuation of signal strength with line length. As further evidence of this, an attempt to fit an exponential decay to the Talk Talk dataset is shown in Figure 8 below and the result is rather poor.

3 Statistical Modelling

3.1 Background

(3.1.1) Talk Talk Technology gathers extensive datasets from individual ADSL connections, including values for line length, capacitance and line speed. The high volumes of data offer the prospect of constructing reliable statistical models relating these variables, but at the same time the datasets are believed to contain a significant minority of values that are unreliable.

(3.1.2) It is therefore useful to have an automated means of identifying and removing points in the datasets that would lead to distortions in any model that was constructed from them. The following paragraphs describe one standard way of doing this. There are several textbooks that can be consulted for further information, for example [6].

(3.1.3) The illustrations that we will use from the Talk Talk datasets look at the relationship between two variables, *e.g.* capacitance (on the y-axis) against line length (on the x-axis). There are two considerations in identifying data points that might distort the model. First is the **leverage**, which describes how far away a point is from the main cluster of data in terms of its x-value. For example, there might be a small number of data with very high line lengths, and these would have high leverage. Second is the **standardised residual**, which describes how far the y-value is from the model estimate. Data points with high residuals might simply be spurious measurements, or they might be an alert to ADSL connections that are faulty or need reprofiling. We describe below how the leverage and standardised residual are defined and used.

3.2 Leverage in regression models

(3.2.1) The general situation when carrying out linear regression is to seek a model of the form

$$Y = X\beta + \epsilon\,,$$

where X is the **design matrix**, containing the explanatory variables, β is a set of parameters to be estimated and $\epsilon \sim N(0, \sigma^2)$ are independent normally distributed errors with some unknown but common variance σ^2.

(3.2.2) The maximum likelihood estimators (equivalent to the least squares estimators when the errors are normally distributed) of the parameters β are

$$\hat{\beta} = (X^T X)^{-1} X^T Y\,. \tag{7}$$

The fitted values are then

$$\hat{Y} = HY\,, \tag{8}$$

where

$$H = X(X^T X)^{-1} X^T \tag{9}$$

is the **hat matrix**. The element h_{ij} reflects the contribution of the measured value Y_j to the fitted value $\hat{Y}_i$. Note from (9) that the hat matrix is idempotent, *i.e.* $H^2 = H$, and also symmetric. The diagonal entries $h_{ii} = \sum_j h_{ij}^2$ (also called 'hat values') reflect the overall contribution of Y_i to the fitted values.

(3.2.3) Informally speaking, **leverage points** are data points that are far away from most other observations in terms of their x-values. They have high influence over the fitted values, corresponding to high values in the h_{ii}. Errors in the corresponding Y_i will lead to errors in the fitted values generally. It can be shown that the average of the hat values h_{ii} is $(m+1)/n$, where m in the number of independent variables in the model and n is the number of data points. In general, the hat values all lie between $1/n$ and 1. It is an accepted rule of thumb that a leverage point is characterised by a hat value at least twice the average.

3.3 Outliers

(3.3.1) The **residuals** $\hat{\epsilon}_i$ are the differences between the observations and the fitted values, namely

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i = (I - H)Y, \tag{10}$$

where I is the $n \times n$ identity matrix. In general, if Σ is the covariance matrix of a set of random variables V then the covariance matrix of a linear transformation AV is $A\Sigma A^T$. From (10) and using the fact that H is idempotent, the covariance matrix of $\hat{\epsilon}$ is $(I - H)\sigma^2$.

(3.3.2) The (unknown) variance σ^2 can be estimated by looking at the **residual sum of squares** $Q^2 = \sum \hat{\epsilon}^2$. Explicitly,

$$S^2 = \frac{Q^2}{n - m} \tag{11}$$

is an unbiased estimator for σ^2 and $(n - m)S^2/\sigma^2$ has a χ^2-distribution with $n - m$ degrees of freedom.

(3.3.3) Outliers in the Y_i can be detected by normalising the residuals using estimates of the variance. We use a method based on the so-called 'externally studentised' residuals, defined for each data point by

$$t_i = \frac{\hat{\epsilon}_i}{\sqrt{S_i^2(1 - h_{ii})}}, \tag{12}$$

where S_i^2 is the estimate (11) for σ^2 with data point i removed from the calculation. In other words, each residual $\hat{\epsilon}$ is normalised by dividing by

an estimate of its variance. It is another rule of thumb that outliers are characterised by values of $|t_i|$ larger than 2. The reason for excluding the data point itself in the estimate S_i^2 is that an outlier would itself distort the variance estimate if it were included.

(3.3.4) To have undue influence on the model fitting, a data point must have high leverage (corresponding to a high hat value h_{ii}) and a high normalised residual t_i. These two requirements can be combined by looking at a single measure called DFFITS_i, defined by

$$\mathrm{DFFITS}_i = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}} \tag{13}$$

The rule of thumb for an influential outlier is that its DFFITS value should exceed

$$2\sqrt{\frac{m+1}{n}}. \tag{14}$$

There are other measures for the automatated identification of outliers, but DFFITS [1] is the one that we chose to apply to the Talk Talk dataset.

4 Statistical Analysis of the Talk Talk Dataset

4.1 Capacitance data

(4.1.1) The dataset contains separate values for the capacitance of each of the two wires to the communication line. These values should be approximately equal but the raw data (Figure 3) shows that this is far from the case. However, once the outliers are removed, the the relationship is much closer to expectation (Figure 4). The residuals for the original data are shown in Figure 5.

(4.1.2) The next stage of the analysis looked at the variation of capacitance with line length, which transmission line theory predicts will be linear. The raw data and also the data after removing outliers and fitting the regression line are shown in Figures 6 and 7, respectively.

4.2 Line speed data

(4.2.1) The third analysis that we undertook was to look at the variation of line speed with both line length and capacitance. The raw data for both of these comparisons are shown in Figures 8 and 9. Both plots are highly scattered, but there seems slightly more structure to the plot of speed against capacitance, and so this was analysed further.

(4.2.2) If the outliers are removed and an exponential fit made for line speed as a function of capacitance, then the result is shown in Figure 10. However,

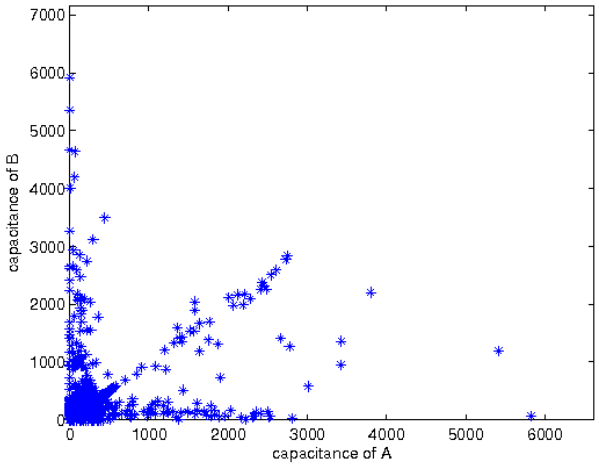

Figure 3: The capacitance of wire B to earth versus the capacitance of wire A to earth before the removal of outliers. All units are in nanofarads.

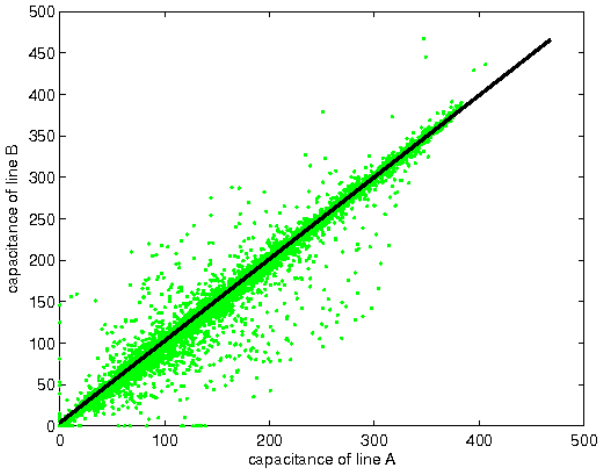

Figure 4: The capacitance of wire B to earth versus the capacitance of wire A to earth, after the removal of outliers, together with the corresponding regression line. All units are in nanofarads.

the fit is not very good and we believe that different functions would give a better fit. This is where the bringing together the electrical modelling of the connection line and the information modelling of the communication channel will provide important guidance in constructing good statistical models. One might also be able to make good use of the techniques that are proposed in [3] for deciding whether datasets display power-law behaviour.

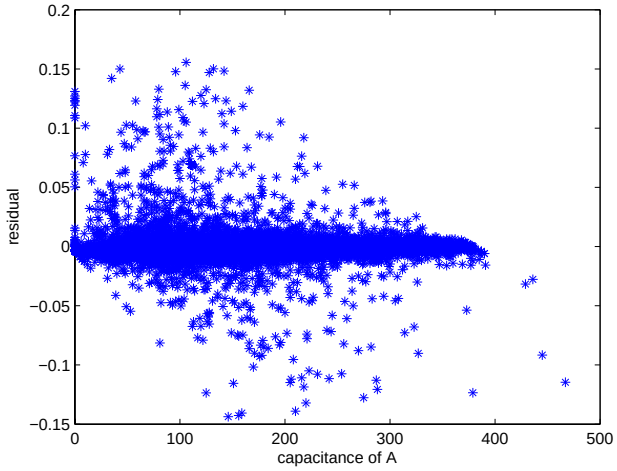

Figure 5: The residual plot for fitting linear regression analysis for the capacitance of wire B to earth versus the capacitance of wire A to earth

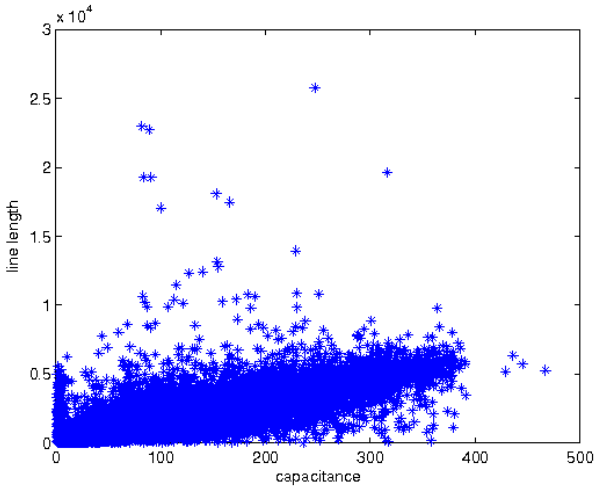

Figure 6: The line length in metres versus capacitance in nano-Farads.

5 Conclusions

5.1 Review

(5.1.1) We have discussed how modelling the behaviour of an ADSL connection involves a combination of (a) electrical modelling of the physical wires, most naturally via a transmission line model, and (b) channel capacity calculations from information theory, in order to relate the supported line

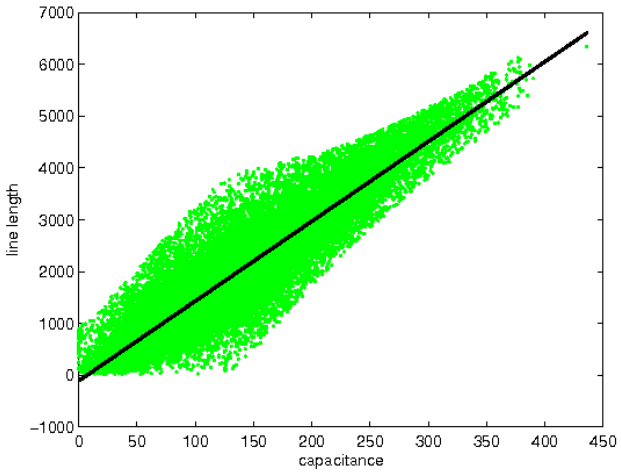

Figure 7: The line length in metres versus capacitance in nanoFarads and the regression line.

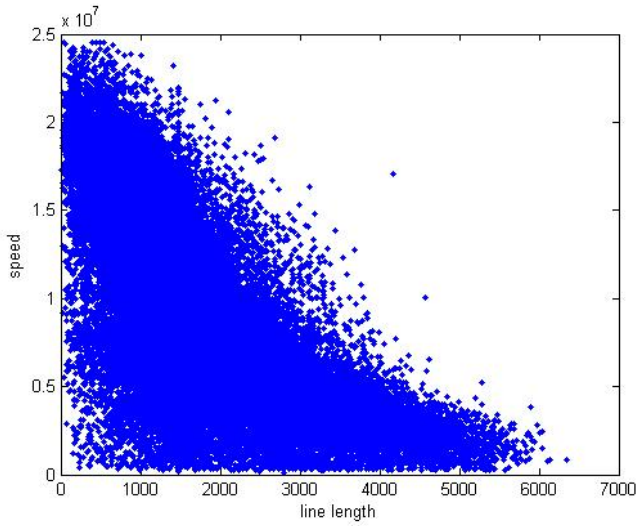

Figure 8: Raw data for line speed (bps) against capacitance (nF).

speeds to the physical properties of the connection. This short report provides only a brief overview, and there are more detailed analyses available in, *e.g.*, [4] and [5].

(5.1.2) We have also discussed how statistical techniques can be used to analyse the datasets that are gathered from ADSL connections. The three main variables in the Talk Talk dataset that we studied are line length, capacitance and line speed. It is important to have a thorough grounding in the physical and information modelling in order to get greatest benefit from

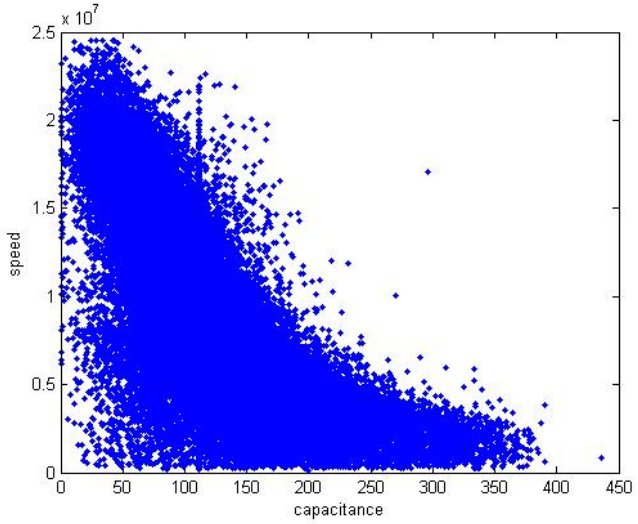

Figure 9: Raw data for line speed (bps) against line length (m).

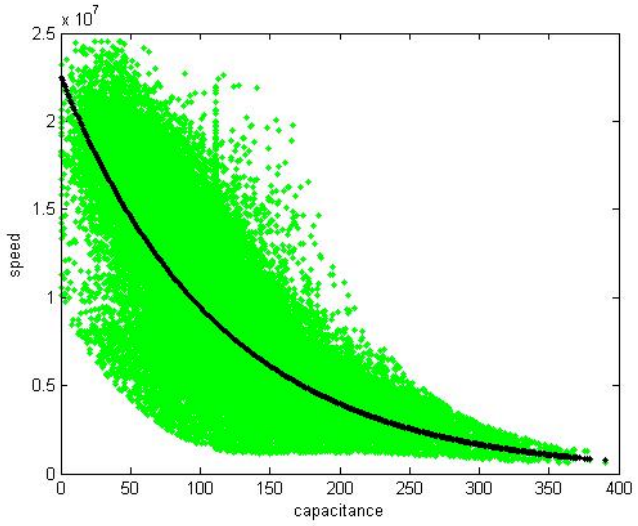

Figure 10: An exponential fit for the line speed (bps) in terms of capacitance (nF) after the removal of outliers.

the statistical analysis.

(5.1.3) In particular, we have described how to automate the identification of unusual or outlying data, which can then either be discarded as spurious or flagged up for further investigation.

5.2 Recommendations

(5.2.1) Our overarching recommendation is that further development of these models should incorporate the three aspects of electrical modelling, information modelling and statistical modelling. There is considerable scope for increasing the systematic statistical analysis of connection line data in order to monitor and predict performance. More specific recommendations are as follows.

- Investigate further the automated detection of outliers in the data, in particular the extent to which the identified data points represent faulty lines and lines on the wrong profile.
- Use more detailed electrical and information modelling to clarify the expected form of the relationships between the measured quantities, which in turn will enable the construction of statistical models that better reflect the true operation of the line.
- Consider whether other quantities could be added to the datasets or whether statistical models could usefully incorporate more than two variables (*e.g.* is there any advantage in building a regression model for line speed in terms of both line length and capacitance?).

(5.2.2) To conclude, we return to the original questions posed in the introduction:

- **Can a better estimate of speed against line length, resistance and capacitance be made?** The answer seems to be yes, through using robust statistical methods, but a more detailed information model is needed to ensure a good fit to observed performance.
- **Is it possible to identify systematically the errors in supplied parameters (*e.g.* line length)?** Using the linear relationship between line length and capacitance seems the best approach, with errors in the line length corresponding to points of high leverage or to outliers.
- **Can we distinguish between lines on the wrong profile and faulty lines?** At present, this is less clear, but perhaps some field experiments could be carried out on the data points that have been flagged up as being outliers. To what extent do outliers correspond to lines that are faulty or on the wrong profile? Once this is known, one could consider the further question of whether faulty lines and lines on the wrong profile could be distinguished from each other by looking more closely at leverage values and standardised residuals.

Bibliography

[1] D. A. Belsley, E. Kuh and R. E. Welsch, *Regression Diagnostics: Identifying Influential Data and Source of Collinearity*, Wiley (1980).

[2] J. M. Cioffi, *A Multicarrier Primer*, available (as of June 2010) at http://www.stanford.edu/group/cioffi/documents/multicarrier.pdf.

[3] A. Clauset, C. R. Shalizi and M. E. J. Newman, *Power-Law Distributions in Empirical Data*, SIAM Review, Vol. 51 pp. 661–703 (2009).

[4] P. Golden, H. Dedieu and K. S. Jacobsen (eds.), *Fundamentals of DSL Technology*, Auerbach Publications (2005).

[5] M. Randelović, A. Atanasković and N. Dončov, *Estimation of ADSL and ADSL2+ Service Quality*, Proceeding of the 8th IEEE International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Services (TELSIKS 2007), 56–59.

[6] T. P. Ryan, *Modern Regression Methods* (2nd Edition), Wiley (2008).