

**MODELLING WITH NEURAL NETS  
THE SERVICE AGREEMENT CASE**

**Pierre Marie Windal  
Nathalie Gouénard  
Christine Oneto**

**This paper describes an operational model used within the scope of a budget process for a warranty extension product. The model is built on a database containing a record of the costs incurred with several thousand vehicles over a period of five years. Each component was modelled in two different ways : The «classic» way, using linear or intrinsically linear models, by means of regression, and traditional distribution functions ; an «alternative» way, based on neural networks. The transition from a calibration achieved by traditional methods to one achieved by means of neural networks does not in any way affect the model itself. On the other hand, the tremendous flexibility of neural networks has significantly improved the model's predictive capabilities, without inasmuch adding to its complexity.**

## INTRODUCTION

A car manufacturer sells not only cars, but also services. A warranty extension, i.e. the extension of the contractual warranty for a given period of time or mileage, is just one example of the services commonly associated with the sale of cars. With this type of agreement, the price is known in advance. The costs, however, are unpredictable and spread out in time. It is vital to find a way of predicting these costs in order to be able to pitch the price of the agreements optimally. If these are sold too cheap, they become a financial drain on the car manufacturer. If they are too expensive, customers will be discouraged from subscribing to them.

Car manufacturers have therefore acquired the tools required for monitoring and predicting the cost of these agreements, first of all by compiling databases of costs, and then by designing models to predict and explain these costs.

The proposed paper describes the forecasting model used by Automobiles Peugeot in its quest for greater predictive awareness of the costs of a typical «Service Agreement» - a warranty extension product.

## THE SERVICE AGREEMENT

The Service Agreement includes two types of service:

1. Cover for events provided for in the manufacturer's maintenance plan: worn components and periodic maintenance.
2. Cover for random breakdowns or faults.

These complementary services extend the manufacturer's contractual warranty for a predefined period and mileage, the current ceilings being 48 months and 120,000 km. The customer selects a distance that matches his requirements – from 40,000 km to 120,000 km in steps of 20,000 km – and the contract expires when that distance has been covered, regardless of the vehicle's age. The actual contract period thus varies depending on the distance covered and the elapsed time, and can be terminated by either of them.

The events occurring under such contracts vary widely from one vehicle (customer) to another.

- Some contracts involve no incidents, while others «collect» them.
- The faults observed, and thus the costs incurred, range from single components to complete systems.
- Repair costs depend on where they are carried out, and this location affects the cost of labour.
- The frequency and nature of the incidents depend on whether they involve a «low-mileage» or a «high-mileage» driver (10,000 km or 30,000 km per year, respectively).
- Incidents occur randomly, apart from some regular features that have been empirically noted (with a significant increase after the contractual warranty expires and during the last months of the warranty period).

Generally, all factors associated with the everyday maintenance of the vehicle can easily be calculated. By contrast, the valuation of faults, which are essentially random, is built up over time by means of a database.

## THE DATABASE

The database lists all the agreements and all the incidents that have incurred reimbursements throughout warranty periods since the Service Agreement began (1992). Precise information is therefore available about the following:

- For each contract (*vehicle base*)
  1. The vehicle's identification: vehicle code, product line, engine.
  2. The initial date of the «manufacturer's» contractual warranty, which marks the start of the agreement.
  3. The agreed distance.
- For each incident (*incident base*)

1. Description: faulty component, whether component was replaced or not.
2. Date of repair.
3. Age of vehicle at time of incident.
4. Odometer reading at time of incident.
5. Cost of repair (parts and labour).

For each product line, the «vehicle» base contains tens of thousands of records, and the «incidents» base contains even more. It therefore lends itself to the activities of data mining and modelling.

### **THE MANUFACTURER'S EXPECTATIONS**

Hitherto, the costs of warranty extension agreements were monitored by ascertaining the cost of past agreements, i.e. those made at least four years earlier, broken down by the year the warranty began, the agreed distance and the type of vehicle. The cost was obtained by dividing the total amount for all incidents, adjusted for inflation, by the total number of agreements.

The partial cost of unexpired agreements was extrapolated to the end of the agreement as a function of the distribution of expenses over time as determined from the expired agreements. If, for example, it was noted that 80% of the expenses of this or that type of vehicle were incurred in the 42nd month, the amount of expenses in the 48th month was obtained by simple linear interpolation. Such calculations were of course refined by technical analysis in order to provide for an empirical adaptation of the pattern of distribution of expenses over time to the developments that were noted or predicted.

#### **Gaps in the exploitation of the data**

A descriptive exploitation of costs suffers from a certain number of drawbacks, e.g.:

1. The dispersion of the unit cost of an agreement as a function of the year of reference, with no indication of whether this dispersion is a trend or merely anecdotal.
2. The merely partial use of the database, since only expired agreements are considered. Cost prices are based on the oldest vehicles.
3. The inability to depart from the existing state of affairs other than by linear interpolation or extrapolation: one single agreement period, five distance cross sections. In particular, the variation of costs over time and over the distance covered during the agreement is unknown.
4. The inconsistencies in historical costs (a cost rising non-monotonically as a function of the agreed distance) associated with sampling problems (insufficient sample) or behaviour (not all the agreed distance being used up by the customer).
5. The inability to estimate the cost of a distance greater than the permitted maximum, or lying between two existing distance cross sections, other than by linear interpolation.

These difficulties can only be overcome by having recourse to modelling, which alone is capable of «smoothing» the details of the database (noisy data), providing a continuous view of costs and simulating situations outside the field of experience (new offers).

#### **The primary objective**

The manufacturer wishes urgently to establish standard costs. These standard costs would of course be based on historical ones, but would not be enslaved by them. They would to some extent provide a simplified but robust picture of the range of costs. Such standard costs would enable the following in particular:

1. Drawing up annual operating accounts for the «end of agreement».
2. Monitoring an annual budget based on standard costs and actual costs in order swiftly to pick up any abnormal deviation.
3. Optimizing rates to take account of both cost prices and the competition.

A knowledge of the standard costs for each product line, agreement period and agreed distance allows a wide range of tariff simulations, making it unnecessary to carry out a global analysis of the agreement portfolio with a constant mix of vehicles. It then becomes possible to put a precise figure on the profitability of an advertising or

promotional campaign targeted at a given type of agreement, vehicle or customer. It is also easy to test hypotheses about the sales mix of agreed durations and distances.

## THE MODEL

### The approach

The key aim of the project is to assess the mean unit cost of a variable period or mileage-based Service Agreement (e.g. 36 months and 100,000 km). The problem thus consists simply of determining an empirical mathematical formula or sets of formulas linking the mean total cost of a Service Agreement to the two parameters which the car manufacturer can control: the duration of the Agreement and the mileage covered.

The main difficulty in doing this is due to the unpredictable and random nature of the serviceable incidents. Though the age of the vehicle may be known with absolute certainty, its mileage will only become known in the event of an incident. This problem initially led the manufacturer to sideline the «mileage» element in the cost prediction model. The global cost was extrapolated from the cumulative cost to date using standard response curves for the family of vehicles under consideration.

At a later stage, the mileage element was expressly acknowledged by performing a probability calculation of the mileage curve of individual vehicles. This calculation provided a medium for expressing the mean unit cost of a Service Agreement in terms of the outcome, summated across the time and mileage variables, of the product of three components:

1. The probability of the vehicle's passing through each of the time/mileage segments under consideration.
2. The probability of its being involved in an incident in each of these segments.
3. The expected value of the cost of the incident.

All the elements of the model, including data processing, parameter calculation, and presentation of results were integrated behind an ergonomic Windows interface.

We describe in the following sections a sequence for obtaining and modelling the data. We begin by establishing the law linking the specific cost of an incident with the time and mileage. We then reconstitute the individual "mileage trajectories" of the vehicles involved in incidents on the basis of their history of incidents and a "mean" trajectory. From this we can deduce the empirical probabilities of transit and incident, which we then need "only" model, putting the whole together to obtain a formula for the mean total cost of the Service Agreement.

## THE DATA

### The mean specific cost

The mean specific cost of an incident  $C(t,k)$  increases with time and mileage. The older a vehicle becomes (in age and mileage), the more it costs. This relationship, however, is valid only on average. On an individual level - that of a particular vehicle - the relationship is not significant because of the great diversity of the components liable to give rise to an incident.

The problem then arises of selecting the correct time and mileage observation window in order to calculate the aggregated data. In fact, if time is selected as the basis, coverage is good for the time variable (*from 1 to 48 months*) but poor for the mileage variable (*from 4,312 km to 48,895 km*). If mileage is selected as the basis, coverage is good for the mileage variable (*from 2,723 to 110,673 km*) but poor for the time variable (*from 25 months to 35 months*). Depending on the observation window - whether time or mileage - a fairly wide variability is obtained not only in the costs but also in the explanatory variables (time and mileage). Now, the weight of the variables depends on this variability.

It has finally been noted that only a balanced mixture of the two observation windows, i.e. the concatenation of an equal number of cross sections of age and mileage, produced statistically significant coefficients with the correct sign under a linear regression of the age and mileage of the vehicle over the mean specific cost.

## Mean and individual mileage trajectories

Calculating the probabilities of passage through the segment  $(t,k)$  requires that the automobile trajectory be known for each vehicle in the sample.

The first step in reconstituting the individual trajectories is to obtain a relationship between the mileage covered and the time. The second step is to reconstruct the individual trajectory by updating the mean trajectory with the vehicle's incident history. Clearly, this updating will be of greater precision where there are numerous incidents spread over time. The updating is achieved by non-linear interpolation in such a way that:

- ... the individual trajectory passes through the mileages recorded at the time of the incident;
- ... between two incidents, the trajectory follows the mean theoretical curve.

Priority is thus given to the recorded mileage, while respecting the mean trajectory. This prevents abnormally low or high mileages distorting the overall trajectory, thanks to the "gravitational" attraction of the theoretical curve.

At the end of this stage, we know the mileage at time  $t$ ,  $t=1, \dots, 48$  for each vehicle in the sample that has been involved in an incident.

## The probability of passage

Next,  $P(t,k)$  is calculated: the empirical probability of a vehicle's passing through the segment  $(t,k)$ . Passing through the segment  $(t,k)$  means that, during the interval  $(t-\Delta t, t+\Delta t)$ , the vehicle will have travelled between  $t-\Delta k$  and  $t+\Delta k$  kilometres. The calculations were performed for 48 intervals of one month and 36 cross sections of 4,000 kilometres. This grid was then reduced to 12 four-month periods and 12 cross sections of 12,000 km in order to smooth the results.

The procedure is simple: for each vehicle, a matrix  $M\{t,k\}$  of 48 rows (*time*) and 36 columns (*km*) is calculated. For each row  $t$ , just one column  $k$  is equal to 1, the others being equal to 0. This column is that which corresponds to the reconstituted mileage  $k$  of the vehicle at time  $t$ . All the matrices  $M\{t,k\}$  are then added to obtain the number of «calls» in the cells  $(t,k)$ . The probability  $P(t,k)$  is obtained by dividing this quantity by the total number of calls. This is the quantity that must then be modelled.

## The probability of incidents

The probability of an incident  $I(t,k)$  is equal to the ratio of two quantities: the number of incidents stated in the cell  $(t,k)$  divided by the number of "useful" passages in the cell  $(t,k)$ . The denominator of this ratio is restricted to the number of useful passages. A useful passage is a cell susceptible of being visited by a vehicle. For a given vehicle, cells that are inaccessible because of age (*the agreement not yet having expired*) or mileage (*the agreed mileage having been exceeded*) are eliminated. The number of incidents stated in the cell  $(t,k)$  is the result of a simple count.

## MODELLING THE EMPIRICAL DATA

In this step the functions  $f$ ,  $g$ ,  $h$  and  $u$  are found such that:

$$\begin{aligned} C(t,k) &= f(t,k), \text{ where } C \text{ is the mean specific cost of an incident recorded at time } t \text{ and mileage } k. \\ P(t,k) &= g(t,k), \text{ where } P \text{ is the probability of passage.} \\ I(t,k) &= h(t,k), \text{ where } I \text{ is the probability of an incident.} \\ T(t) &= u(t), \text{ where } T \text{ is the mean mileage trajectory.} \end{aligned}$$

Each component was modelled in two different ways:

- The «classic» way, using linear or intrinsically linear models, by means of regression, and traditional distribution functions.
- An «alternative» way, based on neural networks.

### «Classic» modelling

### ***Mean mileage trajectory***

The mileage covered in time  $t$  approximately obeys a semi-logarithmic law:  $T(t) = a + b \times \ln(t)$ , where the symbol  $\ln(t)$  indicates the Napierian logarithm of  $t$ .

Graph 1 here

### ***Mean specific cost***

Only over part of the response curve is the mean specific cost a linear function of time and mileage covered. Although mean cost always increases with the age of the vehicle, it is affected differently by mileage depending on whether the vehicle is «young» or «old». For a given distance, «high-mileage» drivers, i.e. those customers who rapidly attain high mileages, in fact tend to suffer less costly incidents than «low-mileage» drivers. This type of interaction can be incorporated into a linear model by adding cross terms (time  $\times$  mileage). It is also necessary, however, to determine the vehicle age at which the relationship between mileage and incident cost tends to be inverted. As a first approximation, we have settled for a strictly linear model:  $C(t,k) = a + b \times t + c \times k$

Graph 2 here

### ***Probability of passage***

For a given four-month period, the probabilities of passage, i.e. the distribution of the mileage covered, follow with remarkable precision a gamma law:  $P(k|t) = \text{Gamma}(\lambda, r)$ , where  $\lambda$  and  $r$  are the two parameters of the relationship. Thus, as many different relationships must be estimated as there are four-month periods, i.e. 12 sets of two parameters.

Graph 3 here

### ***Probability of an incident***

The empirical probabilities of an incident are more uneven than the probabilities of passage. Nevertheless, a certain number of regularities are noted that give the response curve the appearance of a saddle.

Graph 4 here

- **"Start of agreement" effect.** There is a strong probability of an incident during the first four months of application. It then decreases steadily until the "end of agreement" effect in the last four-month period.
- **"End of agreement" effect.** The probability of an incident increases again at the end of the agreement for «balance of whole account».
- **"Low-mileage driver" effect.** Apart from the end-of-agreement effect, the probability of an incident falls with time for "low-mileage drivers", except during the first mileage cross section. In that cross section, the probability increases with the age of the vehicle. This mileage cross section must therefore be subject to a special law.

A semilogarithmic model with one dummy variable allows both of the main effects to be considered: start and end of the agreement. The exponential decrease takes account of the first, while the dummy variable takes account of the second. The model is as follows:

$$\ln[I(t,k)] = a + b \times t + c \times k + d \times D \quad \text{for } k > 1 \text{st cross section}$$

where  $t$  represents the age of the vehicle,  $k$  its mileage and  $D$  is a binary variable equal to 1 at the end of the agreement and 0 elsewhere.

A slightly different model is used for a "low-mileage driver" ( $k=1$ ) - the logistic model:  $\ln[(1-I)/I] = a + b \times t$

where, for the sake of clarity, the letter I denotes the probability  $I(t,k)$ . By construction, this probability is confined to the interval (0,1).

### **Neural modelling**

We note empirically that the laws used in order to formulate the four components of the model (mileage trajectory, mean specific cost, probability of passage and probability of an incident) are nonlinear laws. Moreover, two components require specific sub-models for the optimal definition of certain local phenomena (probability of an incident) or successive estimations to take account of the specific nature of the parameters (probability of passage). The technical difficulty of selecting the most appropriate nonlinear function for the phenomenon studied (log-linear, gamma, logistic) is compounded by the operational difficulty of managing several subsets that vary in their method of calibration.

In addition, although the quality of fit is very good in general, the sub-model of mean specific cost sometimes proves incapable of taking account of the complexity of the response curve. Thus, depending on the vehicles studied, we move from a «satisfactory» ( $R^2 \cong .80$ ) to an «unsatisfactory» level ( $R^2 \cong .50$ ) in the eyes of the end-user. This imprecision in the model is due to its linear character, which by definition is ill suited to taking account of the interactions detected between the age and mileage of the vehicle and, in addition, difficult to integrate simply into an intrinsically linear model.

In our contradictory search for greater simplicity and precision, neural networks quickly became indispensable. A range of methods devised by neurostatisticians since the 1940s (McCulloch & Pitts, 1943), neural networks have been gaining in popularity for some ten years, thanks to the development of fast and high-performance learning algorithms such as retropropagation. They are in use everywhere, including Esomar (Yahiaoui et al., 1997; Stadler & Liehr, 1997). Cf. Chester (1993) for a simple but rigorous introduction to these methods.

### ***General diagram of a neural network***

A neural network consists of a set of small interconnected modules (neurones), each of which receives and transmits information. A network is characterized by three features: activation function, architecture and learning process.

[Graph 5 here](#)

The activation function transforms the input signal ( $X_j$  in the graph above) into the output signal ( $Y_j$ ). Usually, it is a sigmoid function, i.e. shaped like an «S». The architecture defines the number of «layers» in the net and the number of neurones per layer. It also specifies the type of connection linking the neurones to each other (loop or unidirectional flow). In the present case, we have used a network of the Multilayer Perceptron type, illustrated below, for all modules.<sup>i</sup>

[Graph 6 here](#)

This is a network with an input layer having 1 or 2 neurones for the explanatory variables (age and/or mileage of the vehicle depending on the module), an output layer containing a single neurone (cost, mileage trajectory, probabilities of passage and of an incident) and, depending on the phenomenon studied, with 1 to 3 neurones in the hidden layer.

- It is intuitively easy to see the potential of neural networks in the domain of data analysis<sup>ii</sup>. In fact, we note that the weighted sum of the inputs to a neurone defines the left-hand part of the equation of a hyperplane (in two dimensions: that of a straight line) as shown by Graph 5. Finally, with 1 neurone, we have a parametrizable straight line (or, more generally a hyperplane). The use of this straight line as a «parametrizable curve» immediately reminds us of «linear regression», which involves finding the parameters of the straight line most closely fitting the measured points. The difference between linear regression and neural networks is that the latter generally use several neurones, organised into successive layers, for example, which gives the output neurones a capacity of approximating to a phenomenon, however nonlinear it may be (by «stacking» nonlinear functions F). But the basic idea is still the same as in linear regression: we have a parametrizable layer and we seek those parameters (for neural networks: the values of the synaptic weights) that will most closely fit the measured points. The interest of neural networks (in particular layered neural networks) is that they have been shown to constitute «universal approximants», i.e.

that the parametrizable curve they implement can take any form whatsoever (not just a straight line, a spline, etc., as is the case with the classic methods). The nonlinear operation of a multilayer neural network, on the other hand, prevents formal calculation of the parameters, so they are calculated using iterative methods (the best known being the gradient descent function and simulated annealing). These iterative methods for finding parameters in order to optimize a cost function are known by convention as «learning rules».

### ***Taking account of complexity***

After the type of network best adapted to our regression problem has been selected, it remains to specify the number of neurones in the hidden layer. This number determines the plasticity of the network, i.e. its capability for faithful reproduction of the output variable of the network.

For each component in the model, the following table recapitulates the nature and number of variables involved and the number of parameters to be estimated. For the «classic» model, this number (N+1) depends only on the number «N» of explanatory variables. For the «neural» model, the number of parameters to be estimated also depends on the number «n» of neurones in the hidden layer, i.e.  $1+n \times (N+2)$ .

Table 1 here

In the current state of research, in spite of the abundant «theoretical» literature on the subject, no simple solution exists to the problem of choosing the «optimal» number of neurones (Thiria et al., 1997). The choice of the number of neurones reflects the bias/variance dilemma. A good estimator will be characterized by good precision, i.e. little bias, and good stability, i.e. low variance. Now, these two objectives are mutually contradictory: when the number of neurones is increased, the bias is reduced (improving precision), but the variance is increased (reducing robustness). The quality of the estimators cannot but be reduced by adding neurones to a network. However, the complexity of the phenomenon studied may necessitate a minimum number of neurones below which the network will be incapable of reproducing the desired output, whatever the size of the sample and the number of learning cycles. In the present case, the number of input/output variables is constant, but the number of neurones varies from 1 to 3.

Apart from the techniques of validation and resampling (cross, jackknife, bootstrap validation), which are heavy in standard statistics (analytic solutions), and extremely heavy in neural networks (iterative solutions), the practitioner will have to be content with such safeguards as:

- Using software incorporating algorithms that tolerate a certain over-parametrization of the network, by means of techniques of numerical regularization (reducing the noise in the data, pruning, constraints on weights and the objective function, premature stop method etc.).<sup>iii</sup>
- Remembering that the best model is the simplest one. In practical terms, this means it will always be worth resisting the temptation to add neurones to gain precision.
- Stopping the learning process early to prevent the network learning the data by heart.

## **THE RESULTS**

The following table summarizes the various components of the explanatory model of the standard mean total cost of a service agreement with the duration T and agreed mileage K.

Table 2 here

The quality of fit of the model with the empirical data will be illustrated with the aid of a «vehicle+incidents» database containing 144,738 records. This information is aggregated to obtain the data proper to each component: mileage trajectory (N=40 cross sections), mean specific cost (N=12+12 = 24 cross sections), probability of passage (N=12×12=144 cells), probability of an incident (N=144 cells). The aggregated samples are small in size, but the information is much less noisy than that in survey data because each observation is a mean based on several tens or hundreds of vehicles/incidents. The table below shows the percentages of variance explained (R<sup>2</sup>) by each method for each of the four components of the model and two vehicle segments (petrol and diesel engines).

Table 3 here

The quality of fit, in terms of  $R^2$ , is very good for both methods: classic and neural-network. This precision, an indispensable prerequisite for the acceptance of the model, has reassured users as to the validity of «mathematical» developments, which are always considered as an evil, if a necessary one.

The good performance of the classic tools (nonlinear transformations, distribution methods), as compared with the strike force of neural networks, does not detract in any way from the performance of the latter. In fact, the modelling effort is much less with neural networks than with the classic methods. In one case, a parameter (the number of neurones) is adjusted; in the other, the appropriate laws have to be identified and combined (log-linear, logistic, gamma).

The neural network has made it possible to improve appreciably the precision of the mean specific cost, the only linear component in the classic approach. In the case of the petrol engine,  $R^2$  is improved from 0.75 to 0.96. To the extent that this cost is an essential component of the system, this improvement on its own justifies the transition to neural networks.

The choice of the number of neurones is facilitated here by the presence of very clear thresholds in the progression of  $R^2$ . It is clear, for example, that it is necessary to change from 1 to 2 neurones for the calculation of the mean specific cost ( $R^2: 0.59 \Rightarrow 0.95$ ), but that it is pointless to add a third neurone ( $R^2: 0.95 \Rightarrow 0.96$ ). Unfortunately, this will not always be the case.

Neural networks have often been criticized for being grossly over-parametrized statistical tools. It is even asked how such a ridiculous numeric tool (in character recognition, there are sometimes more parameters than observations) can converge with a simple, albeit improved, gradient descent algorithm. Nevertheless, it works well! While, in this case, neural networks in general use two to three times more parameters than the classic method, such is not the case for all modules. In fact, the application of the neural network to the probabilities of passage requires from 9 ( $R^2=0.82$ ) to 17 ( $R^2=0.93$ ) parameters as against 24 for gamma functions (12 times 2 parameters). For equal precision, neural networks are thus not automatically greedier than other methods.

## USE OF THE MODEL

The model provides the standard total cost of a service agreement with the time horizon  $T$  ( $t=1, \dots, T$ ) and the mileage horizon  $K$  ( $k=1, \dots, K$ ), in graphical and tabular form.

[Graph 7 here](#)

This information is systematically broken down by model, engine, country and year. It is clear that a tool of the type described in this article facilitates and enriches the work of the user by allowing him to work more quickly and effectively. It makes simulations possible that were previously avoided for lack of time and the appropriate means. What has been done for the clientele as a whole can in future be done for segments of the clientele without (great) supplementary effort. The possibility of combining the various components of the system by marrying up laws calculated on different databases enables the rapid and statistical specification of standards for recently launched vehicles.

## CONCLUSION

This paper described a model jointly developed by an end-user and a consultant, illustrating the synergy which can result from combining the strong points of both: those of the end-user, which initiates the project and takes the final decisions as to the relevance of the results obtained, and those of the consultant, who implements what is at first nothing more than an intention. This is an operational model, used within the scope of a budget process where the stakes are high. Any errors in making predictions will result in a loss of revenue, if not huge losses.

The paper also illustrates the potential of neural networks as an alternative to traditional modelling. The use of neural networks in this case is amply justified by the need to achieve the greatest degree of accuracy for all components of the system. The transition from a calibration achieved by traditional methods to one achieved by means of neural networks does not in any way affect the model itself. On the other hand, the tremendous flexibility of neural networks has significantly improved the model's predictive capabilities, without inasmuch adding to its complexity.

The specification of neural networks was facilitated here by the nature of the data, which are simultaneously behavioural and aggregated. Thus, the choice of the number of neurones, so important for the model's capacity for generalization, was clear-cut. In the experience of the authors, this is rarely the case; at least, it is never so clearly the case. In general, the development of a neural network goes through a stage of consistent cross validation.

Finally, the quality of the database affects the quality of the final result. Setting up this database requires such efforts that the modelling stage, by comparison, is almost a recreation. In the same way, the time spent detecting and correcting the anomalies, distinctive features, aberrations and other sources of surprise that occurred during the modelling phase greatly exceeds that devoted to the modelling itself. The anticipated gains in terms of the more refined management of tariffs and margins are worth the difficulties encountered. The best resolution of the latter contributes to strengthening the position of the manufacturer in today's automobile environment, more competitive than ever before.

Table 1

**Model parameters**

<b>Components</b>	<b>Variables</b>	<b>Nb Input</b>	<b>Nb Output</b>	<b>Number of neurons</b>	<b>Nb parameters Traditional</b>	<b>Nb parameters Neural net</b>
<b>Average mileage trajectory</b>	Age, mileage	1	1	1	2	4
<b>Expected incident cost</b>	Cost, Age, mileage	2	1	2	3	9
<b>Passing through probability</b>	PTP, Age, mileage	2	1	2/4	2×12	9/17
<b>Incident probability</b>	IP, Age, mileage, expiry time	3+1	1	3	4+2	16

**Legend** : Nb input = number of explanatory or input variables ; Nb output = number of dependent or output variables.

Table 3

**Model performance**  
**Percentage of output explained variance (R<sup>2</sup>)**

<b>Components</b>	<b>Traditional (petrol)</b>	<b>Traditional (gasoil)</b>	<b>Neural net (petrol)</b>	<b>Neural net (gasoil)</b>
<b>Average mileage trajectory</b>	0.96	0.98	n=1 : 0.99	n=1 : 0.98
<b>Expected incident cost</b>	0.75	0.92	n=1 : 0.59	n=1 : 0.85
			n=2 : 0.95	n=2 : 0.96
			n=3 : 0.96	n=3 : 0.98
<b>Passing through probability</b>	0.95	0.97	n=1 : 0.62	n=1 : 0.29
			n=2 : 0.92	n=2 : 0.82
			n=3 : 0.97	n=3 : 0.88
				n=4 : 0.93
<b>Incident probability</b>	0.81	0.90	n=1 : 0.65	n=1 : 0.29
			n=2 : 0.80	n=2 : 0.49
			n=3 : 0.83	n=3 : 0.92
			n=4 : 0.88	n=4 : 0.93

**Legend** : n = number of neurons in the hidden layer.

Table 2

The model

Components	Traditional modelling	Neuronal modelling
Average mileage trajectory	$T(t) = a + b \times \ln(t)$	ML (1)
Individual mileage trajectory	Non-linear interpolation using the vehicle incident history	Idem
Expected incident cost	$C(t,k) = a + b \times t + c \times k$	ML (2)
Passing through probability	$P(k   t) = \text{Gamma}(\lambda, r)$	ML (2/4)
Incident probability	$\ln[I(t,k)] = a + b \times t + c \times k + d \times D$ , pour $k > 1$ $\ln[(1-I)/I] = a + b \times t$ , pour $k=1$	ML (3)
Total standard unit cost	$\text{CMT}(T,K) = \sum_T \sum_K P(k   t) \times I(t,k) \times C(t,k)$	idem

Legend : t=age of vehicle, k=mileage, D=dummy variable for agreement expiry time, ML(n) = multilayer net with one hidden layer of n neurons, parameters to be estimated: a, b, c, d,  $\lambda$ , r and  $1+n \times (N+2)$  parameters for the neural net.

Figure 1

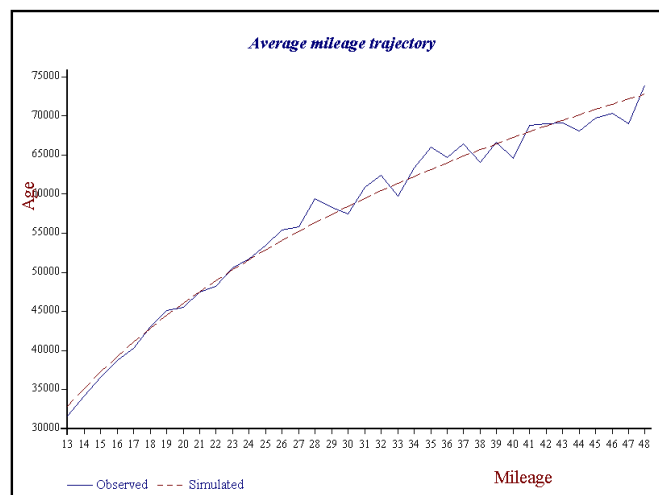


Figure 2

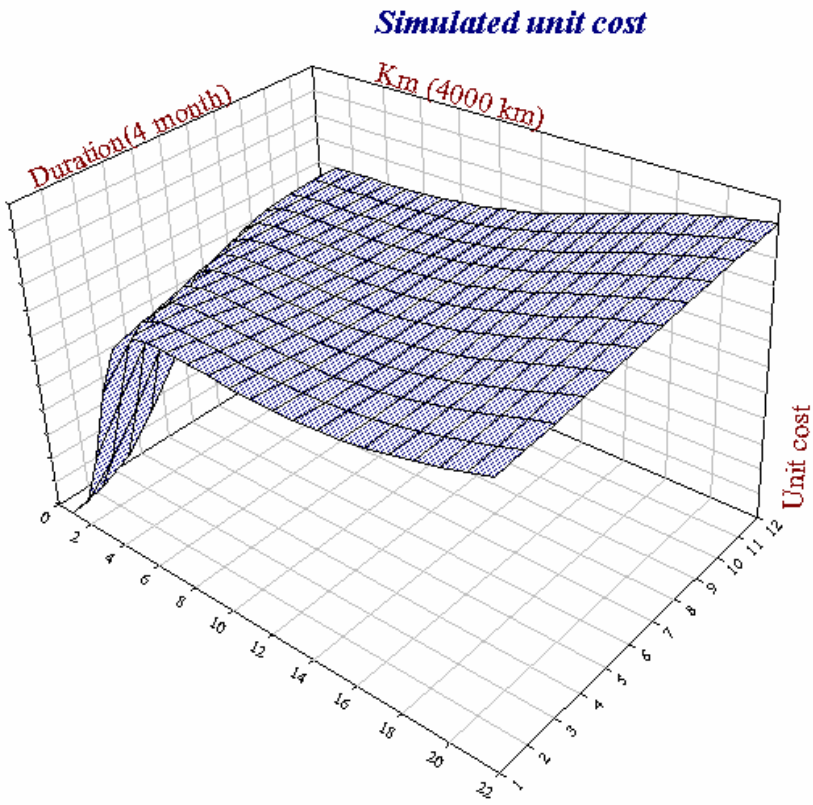


Figure 3

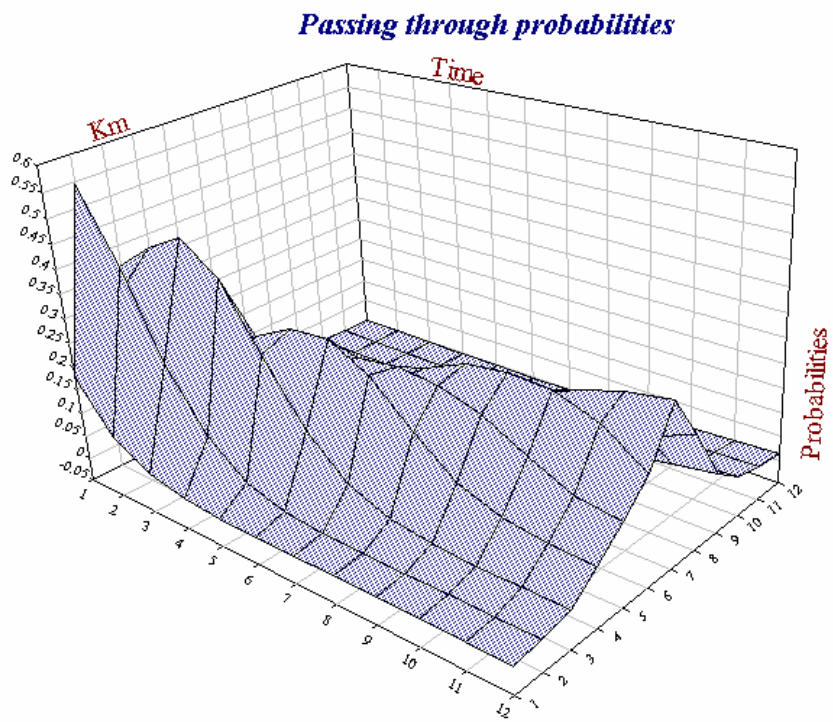


Figure 4

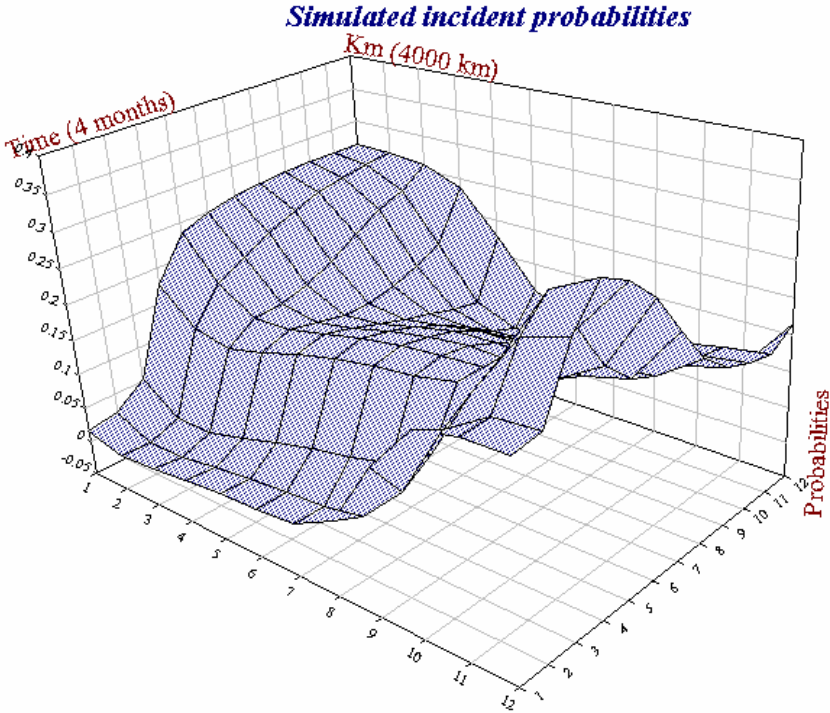
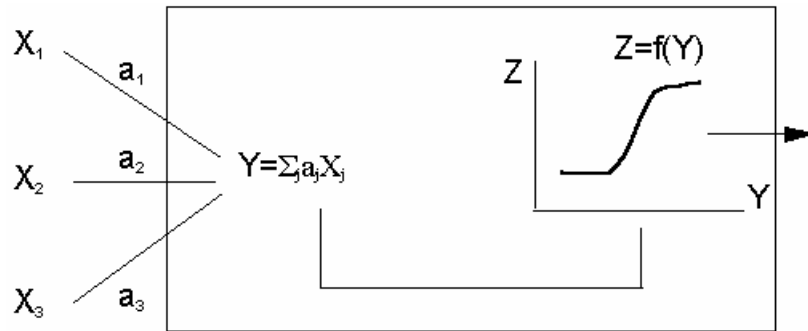


Figure 5

The neuron mechanism



Adapted from Chester M. (1993)

Figure 6

A multilayer neural network

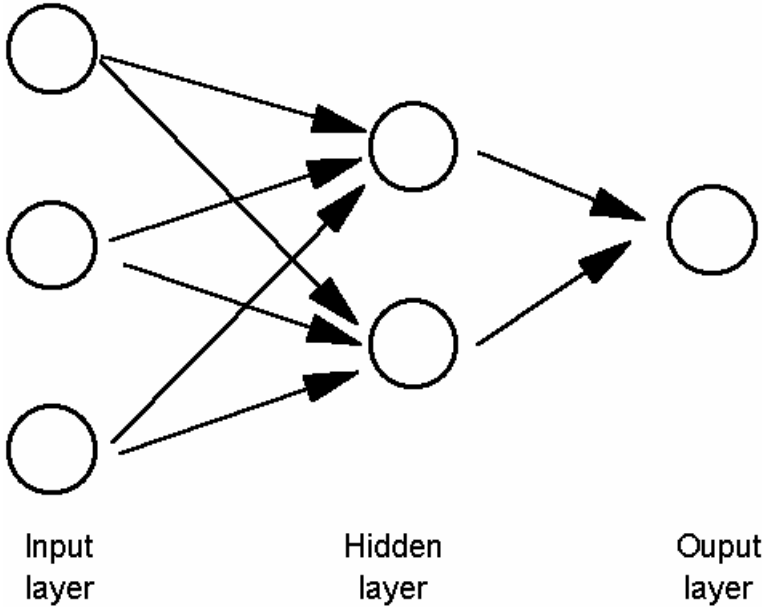
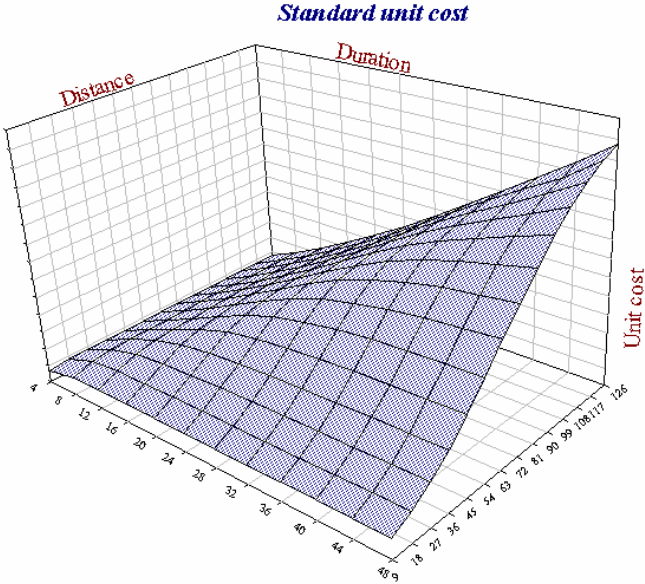


Figure 7







---

## FOOTNOTES

<sup>i</sup> Other types of network would probably enable an equivalent result to be obtained (e.g. Radial basis functions Network, Cascade Correlation). The Multilayer Perceptron was selected for two reasons, one theoretical, the other practical. It is well suited to the problem posed (regression); we had a very high-performance version of this type of network at our disposal.

<sup>ii</sup> Adapted from G. Yahiaoui, private correspondance.

<sup>iii</sup> The Mensor software, developed by Quadrature and incorporating neural networks designed by Nexyad, provides this type of safeguard.

## REFERENCES

Chester M., 1993, *Neural Networks: A Tutorial*, PTR Prentice Hall, Englewood Cliffs, NJ.

McCulloch W. and Pitts W., 1943, A logical calculus of the ideas immanent in neural nets, *Bulletin of mathematical biophysics*, vol. 5, pp 115-133.

Stadtler K. and Liehr T., 1997, Using neural networks to put customer satisfaction data into action, *Proceedings of the 50<sup>th</sup> Esomar marketing research congress*, Esomar, Amsterdam, pp 637-661.

Thiria S., Lechevallier Y., Gascuel O. and Canu S., 1997, *Statistique et méthodes neuronales*, Dunod, Paris.

Yahiaoui G., Da Silva Dias P. and de Saint Blancard M., 1997, Customer segmentation for the automobile market : the use of artificial neural networks, *Proceedings of the 50<sup>th</sup> Esomar marketing research congress*, Esomar, Amsterdam, pp 663-680.

## THE AUTHORS

Pierre Marie Windal is managing director of Quadrature, France.

Nathalie Gouénard is planning analyst at Automobiles Peugeot, Spare Parts and Services Division, France.

Christine Oneto is marketing and research analyst at Automobiles Peugeot, Commercial Methods, France.