S³RI Southampton Statistical Sciences Research Institute
University of Southampton

Estimating Vehicle Excise Duty Evasion Review: Part 1

Pedro Luis do Nascimento Silva
August 2007

Executive Summary

Overall, our initial assessment is that the methodology used by the Department for Transport to estimate the level of vehicle excise duty evasion appears to be sound. It relies on some strong assumptions, namely:

a) the cost of vehicle excise duty is broadly constant within each vehicle tax class;
b) the observed sample of vehicles sighted in the Roadside Traffic Observation Survey is a simple random sample with replacement of the registered vehicles;
c) the relative mileage of vehicles sighted is proportional to their numbers of sightings.

These assumptions are essential for the development of the estimation approach utilized, but appear to be justified given the present survey scenario.

The methodology relies on fitting a Negative Binomial model to observations of repeat vehicle sightings obtained from the Roadside Traffic Observation Survey. The model follows through if the assumptions are made and there is no error there. No mistakes have been found in the calculation of the estimates themselves. However, there are other methodological aspects of the estimation process which have not been included in this initial assessment, and which should therefore be considered for further investigation.

As is standard statistical practice, each estimate given in the publication is presented with a corresponding confidence interval to represent the statistical uncertainty that naturally arises from the sampling process. While important for all estimates, these intervals are particularly relevant when considering the financial estimates of revenue loss through vehicle excise duty evasion.

There are, however, a number of recommendations that should be implemented in order to improve the estimation process. These are as follows.

1. To fit an alternative statistical model to estimate relative mileage when the Negative Binomial is found to be inadequate

   Whilst suitable for most tax classes, the Negative Binomial model was found to be inadequate for the distributions of repeat sightings of 'buses' and 'other' vehicles. This issue occurs when the sample variance is found to be smaller than the sample mean. The model fitting procedures adopted should therefore be improved in order to detect cases which require that a different model is fitted to the data. In both of the vehicle types outlined above, the Poisson model would have provided a suitable alternative but it should be noted that, in practical terms, both the Poisson and Negative Binomial models produce identical estimates of evasion in stock.

2. To use Maximum Likelihood estimation in place of Method of Moments to calculate the parameters of the fitted distribution

   The models currently used are fitted using Method of Moments estimators for the two relevant parameters of the Negative Binomial distribution. While this is an acceptable method, it could be improved upon by using the Maximum Likelihood method to fit the parameters of the chosen models. However, again it should be noted that the estimates for the mean parameter of the model under both Maximum Likelihood and Method of Moments are identical and, therefore, the choice of method does not affect the estimates of evasion in stock.

3. To limit the maximum number of repeat sightings that are considered when fitting the chosen model

   The distribution of the number of repeat sightings of vehicles observed in the survey was found to be quite skewed for some tax classes. This issue was confirmed by the calculation of Chi-square statistics which showed a poor model fit for some tax classes. The model fitting process could therefore be improved upon by limiting the maximum number of repeat sightings considered when fitting the chosen model for each tax class. While this adjustment would increase the statistical robustness of the model fitting procedures, testing has shown that it is unlikely to have a significant impact on the final estimates of evasion in stock.

4. To consider alternative methods to deal with the issue of using both weighted and un-weighted approaches in the model

   The estimation of evasion in traffic takes account of the weighting of traffic in roads of different types in order to reflect different volumes of traffic. However, the estimation of the adjustment factors used to convert evasion in traffic into evasion in stock does not take account of this weighting. Producing both estimates either using weights all the way, or not using weights at all, would lead to a more coherent use of the survey data. We recognise that it would be somewhat complex to produce weighted estimates for the adjustment factors used to convert evasion in traffic into evasion in stock. Nevertheless, the Department should consider alternative approaches of using either non-weighted or weighted estimates throughout the whole estimation process.

5. To revisit underlying assumptions which cannot be verified from the survey data itself

   One of the most important assumptions in the model is that the average number of sightings of a given vehicle is proportional to its mileage. This hypothesis is not testable from the survey data itself because the mileage of individual vehicles is not directly observed through the survey process. However, the first time that this working assumption was adopted - see §4 in Appendix C of (Department of Transport, 1984) – a postal survey of the keepers of heavy goods vehicles was used to test the adequacy of this hypothesis. Given that this research was carried out some time ago and for a limited sample of vehicles in a single tax class, the Department for Transport should investigate whether alternative data sources exist, or could be obtained, which could be used to re-examine the validity of this crucial assumption.

6. To improve the existing documentation regarding the survey and estimation procedures

   There is little documentation or desk instructions currently available to describe the survey and estimation processes. This is a weakness and should be targeted for improvement, particularly for the benefit of new staff working on the survey and its outcomes.

In addition to these recommendations, the following areas, while outside the scope of this initial report, warrant further investigation:

A. The sample design for the roadside traffic observation survey and whether it remains suitable for the purpose of estimating vehicle excise duty evasion;

B. The methods used to weight the roadside survey results in estimating evasion in traffic, and how the survey data are aggregated into different subgroups;
C. Whether alternative estimates of vehicle excise duty evasion can be developed from the roadside survey results;
D. The methods currently used for precision estimation and whether they can be further improved.

1 Introduction

This report is the first part of a review of the approach currently used by the Department for Transport (DfT) to estimate "the stock of evading vehicles" and the "amount of vehicle excise duty evasion". This review was requested by the DfT via the Office for National Statistics (ONS), and is provided under the terms of the "Contract extension for the provision of methodological research and analysis" maintained between the ONS and the University of Southampton.

Following correspondence with Anthony Boucher of DfT and a meeting on April 10 when the key aspects of the project were discussed with him and Drew Hird (also from DfT), it was decided that the main focus of the review of the current methodology should be "the use of the negative binomial distribution to estimate the stock of evading vehicles". It is also of interest to consider other aspects of the methodology, as is described in (Boucher and Hird, 2007), but these might be considered in a second phase, in view of the requested timetable of the project: an initial summary feedback was required by the end of May. It was also requested that the project helped the DfT to document the current estimation approach such that all its steps are easily reproducible.

In section 2 the key targets of inference are defined. In section 3 the current estimation approach is reviewed, highlighting its dependence on some working assumptions, providing some reasoning to explain its development, and reviewing its key aspects. Section 4 identifies the issues that need to be addressed for completing the review, and lists some ideas for research on aspects where some improvement over the current approach may be possible.

2 Key concepts and target parameters

Before trying to evaluate any estimation procedure, it is essential to define what it is trying to estimate. This corresponds to establishing the conceptual framework for the problem, identifying the relevant reference population, and defining the relevant target parameters. This is the goal of this section.

Vehicle excise duty evasion is estimated periodically by the DfT for publication in the Transport Statistics Bulletin (Department for Transport, 2007) as National Statistics. The

main goals of these National Statistics on vehicle excise duty evasion are to estimate both the number (stock) and the amount of tax due (not paid) corresponding to *evading vehicles* in the UK. These are vehicles which do not have an up-to-date tax disc (are *unlicensed*) and are *in use*. A vehicle is considered to be *in use* if it is driven in a public road anywhere in Great Britain or Northern Ireland during the period of observation. Note that this definition of *evading* vehicles depends on two conditions: the vehicles must be *unlicensed and in use*.

Vehicle excise duty (VED) is collected by the Driver and Vehicle Licensing Agency (DVLA) for vehicles which are *registered* in Great Britain, and by the Driver and Vehicle Licensing Northern Ireland (DVLNI) for vehicles registered in Northern Ireland. From now on, we refer to these Vehicle Licensing Agencies simply as VLA. For this reason, the *reference population* of vehicles, denoted *P*, is defined as:

> All vehicles registered as active within the country's vehicle licensing agency (VLA) database at the time of the survey.

Note that vehicles not registered in Great Britain or Northern Ireland are not included. For example, if a vehicle is registered in France and is driven in Great Britain during the time of the survey, it will not be part of the reference population, and correspondingly, will not be considered during the estimation of vehicle and tax evasion. Also, registered vehicles which do not have an active registration (e.g. registered as scrapped or exported permanently) would not be included.

For each vehicle registered in the VLA database, there are six main fields (variables) which are relevant for the purposes of the vehicle evasion estimation (see table 1). Note that some of these fields may actually represent derived variables, such as the VED amount paid or due, which depend on other fields of the database (such as engine capacity). For the purposes of this report, not all the detail is needed and the definitions provided in table 1 are sufficient.

Later we shall see that there is a seventh category for vehicle taxation class: Unknown. This corresponds to a small fraction of records for which the VLA database does not contain a valid code / category, but which are retained in parts of the analysis.

Table 1 – Variables retrieved from VLA database for each registered vehicle

Variable name	Variable description
Registration mark	the vehicle's number plate, which serves as a unique identification code for each vehicle
Tax class	the vehicle's taxation class (see table 2 below for the possible categories of this field)
VED amount paid or due	amount of vehicle excise duty (VED or vehicle tax) paid or due for each vehicle, considering its tax class and any other relevant factors (engine size, gross weight, CO2 emissions, number of axles, etc.), converted to an annual reference period, in pounds
License status	the status regarding the tax disc or whether the vehicle is properly licensed at the time of the survey (two possible categories: either licensed or unlicensed)
Year of first registration	year when the vehicle was first registered with the VLA
Type of keeper	type of registered keeper, with three categories: private, company, or between keepers

Table 2 – Possible categories for vehicle tax class

h	**Code**	**Vehicle taxation classes**
1	PLG	Private and light goods
2	GOODS	Goods vehicles
3	MCYCLES	Motorcycles
4	BUS	Buses
5	EXEMPT	Exempt
6	OTHER	Other

In order to define the target parameters, we need to introduce some notation. Let x_k denote the indicator variable taking value 1 if the *license status* of vehicle k is *unlicensed*, and 0 otherwise, for $k \in P$. Let y_k denote the indicator variable taking value 1 if vehicle k is *in use*, and 0 otherwise, for $k \in P$.

Hence the *number of evading (i.e., unlicensed and in use) vehicles* can be defined as:

$$U = \sum_{k \in P} x_k y_k = \sum_{k \in P} u_k \qquad (1)$$

where $u_k = x_k\, y_k$ is the variable indicating that vehicle k is *evading*, $k \in P$.

It is also relevant to define the domains corresponding to the different tax classes. Let P_h denote the set of registered vehicles belonging to tax class h, h=1,…,6. The corresponding population totals for the indicator variable u_k within tax classes are given by:

$$U_h = \sum_{k \in P_h} u_k \quad \text{for } h = 1,...,6\,. \tag{2}$$

The population totals U and U_h represent the *stock of evading vehicles* overall and by tax classes, respectively. Note that $U = \sum_{h=1}^{6} U_h$. The population means of the indicator variable u_k represent the *proportions of evasion in stocks*, defined as:

$$\overline{U} = \sum_{k \in P} u_k / N = U / N \tag{3}$$

$$\overline{U}_h = \sum_{k \in P_h} u_k / N_h = U_h / N_h \qquad \text{for } h=1,\ldots,6 \tag{4}$$

where N_h is the number of registered vehicles in tax class h, and $N = \sum_{h=1}^{6} N_h$ is the total number of registered vehicles. These proportions are referred to as "evasion in stock (%)" in the main publication containing survey results – see (Department for Transport, 2007).

Let t_k denote the variable *VED amount paid or due* by vehicle k, $k \in P$. Then the total amount of vehicle excise duty for all vehicles is

$$T = \sum_{k \in P} t_k \tag{5}$$

and the corresponding amount owed by all *evading* vehicles, namely the *total vehicle excise duty evasion* (in monetary terms) is given by

$$Z = \sum_{k \in P} u_k t_k = \sum_{k \in P} z_k \tag{6}$$

where $z_k = u_k\, t_k$ for $k \in P$. The corresponding totals for each tax class are defined as:

$$T_h = \sum_{k \in P_h} t_k \quad \text{for } h = 1,...,6 \tag{7}$$

$$Z_h = \sum_{k \in P_h} u_k t_k \quad \text{for } h = 1,...,6 \tag{8}$$

and they satisfy $T = \sum_{h=1}^{6} T_h$ and $Z = \sum_{h=1}^{6} Z_h$.

3 Current estimation approach

3.1 General form of estimators

Given the data in the VLA database, the values of the indicator variables x_k and of the tax values t_k could be available for all *registered* vehicles. Hence most of the information required for calculating the *total vehicle excise duty evasion* (in monetary terms) is available. There is only a crucial bit of information missing: the indicator of whether the vehicle is in use or not (our y_k, and consequently, the u_k). This is the reason why a Roadside Traffic Observation Survey (RTOS) is needed: to obtain the indicators (y_k) that vehicles are *in use*, which are then combined with the indicators that they are unlicensed (x_k) to obtain the evasion indicators (u_k). Since these indicators are only going to be obtained for a sample of vehicles, the target parameters (*U, Z,* and U_h, Z_h for h=1,…,6) can only be estimated.

The *total vehicle excise duty evasion* (in monetary terms) Z is currently estimated using an estimator of the form:

$$\hat{Z} = \sum_{h=1}^{6} \hat{U}_h \times \overline{T}_h \qquad (9)$$

where $\hat{U}_h$ is an estimator of the U_h, number of vehicles evading in tax class h, and $\overline{T}_h = T_h / N_h$ is the average value of tax paid or due for vehicles in tax class h, obtained from the VLA database. The tax evasion by tax class is estimated by $\hat{Z}_h = \hat{U}_h \times \overline{T}_h$ for h=1,…,6.

The estimators ($\hat{U}_h$) of the numbers of vehicles evading in each tax class are obtained using one of two alternative approaches, which are reviewed in the sequence of this report. In addition to their role in estimating the VED tax evasion, these estimates are of interest in themselves (*evasion in stock*), and are published together with the tax evasion estimates – see (Department for Transport, 2007).

The form of the above estimator can be justified by the following approximation:

$$Z = \sum_{h=1}^{6} Z_h = \sum_{h=1}^{6} \sum_{k \in P_h} u_k t_k \cong \sum_{h=1}^{6} \sum_{k \in P_h} u_k \overline{T}_h = \sum_{h=1}^{6} U_h \overline{T}_h \qquad (10)$$

where the approximation is good only if the vehicle excise duty values t_k vary little for evading vehicles *within* each tax class. Under this approach, the crucial element of the problem is then the estimation of the *number of evading vehicles by tax class* (U_h), because

the average tax values per tax class $\overline{T}_h$ are obtained free of sampling error from the VLA database.

Hence the first working assumption behind the current approach to estimate vehicle excise duty evasion is:

H$_1$) $t_k \cong \overline{T}_h$ for $k \in P_h$ and $h = 1,\ldots,6$.

Note that this assumption is not required to justify the estimators utilized for the evasion in stocks (U or U_h) to be discussed in the sequence.

To estimate how many evading vehicles there are in each tax class, a Roadside Traffic Observation Survey (RTOS) is carried out in a sample of 236 sites across Great Britain, plus 20 sites in Northern Ireland, totalling 256 observation sites (Department for Transport, 2007). At each of these sites and for a sample of time periods during a specified survey month, the survey records (either electronically or by hand) the registration marks (number plates) of all passing vehicles. Data collection for the 2006 edition of the RTOS took place between 5 June 2006 and 2 July 2006.

The selection of sites for observation in Great Britain was stratified by two key variables: geographical location and road type. Table 3 below presents a summary description of the stratification and number of sites sampled in each stratum for the RTOS in Great Britain.

Table 3 – Stratification and allocation of the sample of observation sites – RTOS 2006

Type of road	**Geographical location**			
	Greater London		**Other areas**	
	Substrata	**Number of sampled sites**	**Substrata**	**Number of sampled sites**
Motorways	None	4	10 GOR	20
Built-up major (A) roads	None	4	49 PFA	49
Non-built-up major (A) roads	None	4	49 PFA	49
Built-up minor roads	None	4	49 PFA	49
Non-built-up minor roads	None	4	49 PFA	49
Total		**20**		**216**

GOR = Government Office Region
PFA = Police Force Area

The registration marks observed in each country are then matched against the corresponding VLA database. For each valid recorded registration mark the match exercise determines whether the vehicle carries an up-to-date tax disc (*licensed vehicles,* x_k=0) or not (*unlicensed vehicles,* x_k=1). Obviously vehicles spotted in the roadside survey are all *in use* (i.e. y_k=1). Hence we also have the corresponding evasion indicators ($u_k = x_k\, y_k$). The matching exercise also retrieves all the additional information about each vehicle sighted, such as its registered tax class (h), year of first registration, type of keeper, etc. All estimates of vehicle evasion and vehicle excise duty evasion are based on the records obtained from these vehicle sightings after matching to the VLA database.

3.2 "Crude" estimates of evasion

Before considering possible improvements to the current approach for estimation of vehicle and VED evasion, it is important to describe the current methodology in detail, since the available documentation is scarce.

The target parameters U and Z of the Vehicle Excise Duty Evasion defined in (1) and (6) are population totals related to the concept of *evasion in stock*. If we had a standard probability sample of the *registered* vehicles and could collect data on the variables required (namely, x_k, y_k, u_k and t_k) then estimating the population totals would be a straightforward exercise.

However, the survey carried out to collect data about VED evasion is based on *observing traffic* in a fixed sample of locations scattered around the country. In addition, the Roadside Traffic Observation Survey (RTOS) is not a standard probability sample, because although it generates data on a sample of vehicles, these were not selected directly from the frame provided by the VLA vehicle database. Instead, this survey collected sightings of registration marks in 256 sites, sampled "according to methods established in previous surveys" – see (Department for Transport, 2007) p.12 – and then performed reverse matching of the sighted vehicle registration marks with the VLA database.

One consequence of this survey approach is that the same vehicle may have *multiple sightings* during the survey period. In the 2006 edition of the RTOS, the total number of sightings of *valid* registration marks was 1,268,633, corresponding to 1,100,338 registered vehicles (see table 4 for a breakdown by tax class).

Table 4 – Numbers of vehicles and sightings by tax class and licensing status

Tax Class	Licensed		Unlicensed		All	
	Vehicles	Traffic	Vehicles	Traffic	Vehicles	Traffic
PLG	981,128	1,129,437	18,481	19,782	999,609	1,149,219
Goods	44,869	53,689	707	762	45,576	54,451
Motorcycles	8,239	8,940	1,427	1,465	9,666	10,405
Buses	6,120	8,792	36	39	6,156	8,831
Exempt	35,666	41,587	892	965	36,558	42,552
Other	1,450	1,668	140	144	1,590	1,812
Unknown	1,056	1,214	127	149	1,183	1,363
All	1,078,528	1,245,327	21,810	23,306	1,100,338	1,268,633

The total number of vehicle sightings was larger (1,337,621) but a portion of these did not match records in the VLA database, and hence the corresponding data are not used for the evasion estimation. Very limited information is available for these sightings: site where observed, day of the week, time of day, and a simple categorization of vehicle tax class: Private and light goods, Goods, and Other. A simple descriptive analysis of the data on all vehicle sightings is provided in Tables 9 and 10 in Appendix A of (Department for Transport, 2007). Here the sightings resulting in invalid registration marks are not considered for any of the subsequent analyses.

Ignoring the fact that multiple sightings for the same vehicle may be a problem, the sample data can be used to obtain naïve estimates of the target parameters. Basically, the RTOS provides a (very large) sample of sightings of vehicles which are in use. An estimator of the proportion of vehicles evading can be obtained simply by the ratio of the number of sightings corresponding to unlicensed vehicles (23,306), divided by the total number of sightings (1,268,633), leading to an estimated 1.84% evasion rate. The figure quoted in §2.2 (Department for Transport, 2007) for the crude estimate of evasion was given as the ratio of 23,300 sightings of evading vehicles, over 1,268,600 sightings of valid license plates, resulting in “1.8 per cent evasion based on unweighted observations”.

To describe this estimator formally, denote by r_k the *number of sightings* of a vehicle k during the survey observation period. Then the estimator described as 'crude measure of evasion' in §2.2 (Department for Transport, 2007) is given by:

$$\bar{r} = \sum_{k \in s} r_k \, u_k \Big/ \sum_{k \in s} r_k \qquad (11)$$

where s is the set of all sampled *vehicles*.

For any domain of interest, an estimator similar to (11) can be defined simply by computing the sums in the numerator and denominator only for vehicles in the corresponding domain. For example, crude evasion rates by tax class would be estimated by

$$\bar{r}_h = \sum_{k \in s_h} r_k \, u_k \Big/ \sum_{k \in s_h} r_k \qquad (12)$$

where s_h is the set of vehicles sampled in tax class h.

In fact, the so-called crude estimates of evasion based on this approach are published for the total of Great Britain or Northern Ireland (no further regional breakdown), and for some vehicle tax classes by year of first registration or type of keeper of the vehicles – see Tables 11 to 13 in the Appendix B of (Department for Transport, 2007). The numbers of sightings appearing in those tables do not match the figures in Table 4, possibly because some sightings were included in our Table 4 but not in the published tables (they might have some of the fields used to classify sightings missing – e.g. the year of first registration).

Some remarks about this method of estimation are needed. First, this corresponds to an estimator called "ordered sample mean, repeated elements included" obtained under *Simple Random Sampling With Replacement* (*SRSWR*) of registered vehicles from the VLA database – see eq. 3.3.21 of (Särndal, Swensson and Wretman, 1992). To verify this, think about the various sightings as *repeated independent draws of vehicles* from the VLA database, each draw giving any registered vehicle equal probabilities of observation. Then let m ($= \sum_{k \in s} r_k$ in the notation used here) denote the *total number of sightings*, and think of the outcomes of each of the draws (sightings) as either 1 if the corresponding vehicle is unlicensed and 0 otherwise (the sum of these would just be equal to $= \sum_{k \in s} r_k \, u_k$ in the notation defined here).

Hence the assumption required to justify the estimator (11) would be:

> H_2) The observed sample of vehicle sightings resulting from the RTOS is a simple random sample with replacement (SRSWR) of registered vehicles from the population of registered vehicles *P*.

Following result 3.3.4 in (Särndal et al., 1992), a corresponding estimator for the stock of evading vehicles could be obtained simply multiplying $\bar{r}$ by the total number of vehicles in the VLA database, namely:

$$\hat{U}^1 = N \times \bar{r} \qquad (13)$$

The value of this estimate for 2006 would be 35,534 × (23,306 / 1,268,633) = 653 thousand vehicles. The figure of 35,334 thousand registered vehicles was obtained by adding up the total numbers of evading and licensed stock from table 5 in (Department for Transport, 2007). Applying a similar argument leads to estimates for the evading stock in each tax class (results presented in Table 5 below) given by:

$$\hat{U}_h^1 = N_h \times \bar{r}_h \qquad (14)$$

Table 5 – "Crude" estimates of evasion

Tax Class	Crude evasion rate	Registered vehicles (000s)	Crude evasion in stock (000s)
PLG	1.7%	30,928	532
Goods	1.4%	455	6
Motorcycles	14.1%	1,835	258
Buses	0.4%	114	1
Exempt	2.3%	2,081	47
Other	7.9%	121	10
Unknown	10.9%		
All	1.8%	35,534	653

Source: estimates obtained using estimators (13) – for all vehicles – and (14) for vehicles in each tax class, applied to vehicle sighting counts provided in table 4 and numbers of registered vehicles from table 5 in (Department for Transport, 2007).

Although the reasoning used to justify the estimators (13) and (14) is the same, the outcome is not the same. If the estimates for the evading stock in each tax class are added together, the resulting estimator

$$\hat{U}^2 = \sum_{h=1}^{6} \hat{U}_h^1 = \sum_{h=1}^{6} N_h \times \bar{r}_h \qquad (15)$$

leads to an estimate of 854 thousand evading vehicles, much larger than the estimate obtained using (13) – 653 thousand vehicles.

The estimator $\hat{U}^2$ is a *poststratified estimator*, which used auxiliary information on the total numbers of registered vehicles in each tax class (N_h). With this alternative poststratified estimator, the overall evasion rate would be estimated as $\hat{U}^2 / N = \left(\sum_{h=1}^{6} N_h \times \bar{r}_h\right) / N = 854 / 35{,}534 = 2.4\%$.

The estimator in (13) is in general, less precise than the estimator in (15). So even if the data on sightings were to be used to estimate evasion in stock directly, estimator (15) should be preferred. In addition to being more precise, it avoids the uncomfortable fact that the sum of the estimates of evasion in stock by tax class obtained using (14) does not add up to the evasion in stock estimated using (13), although both estimators follow from the same guiding principle: estimate the evasion rate and multiply by the known population size.

Another reason to avoid (13) is the following. All the estimators discussed so far could be justified under simple random sampling with replacement selection of vehicles. Under the same assumed sample design, the discussion in sections 3.8.1 and 3.8.2 of (Särndal et al., 1992) suggests that another kind of estimator is more precise (i.e., has smaller variance) than the estimators considering the repeated sightings. Ignoring the multiple sightings of each vehicle leads to an alternative estimator of the overall *evasion rate* for Great Britain given simply by:

$$\bar{u} = \frac{1}{n} \sum_{k \in s} u_k \qquad (16)$$

where $n = \sum_{h=1}^{6} n_h$ is the total number of sampled registered vehicles, n_h is the number of registered vehicles sampled in tax class h. The value of this estimate, computed using the counts in table 4, is 21,810 / 1,100,338 = 1.98%.

To estimate the total stock of evading vehicles, the corresponding estimator would be

$$\hat{U}^3 = N \times \bar{u} \qquad (17)$$

leading to an estimated evading stock of 704 thousand vehicles.

Again we could apply a similar approach to estimate the proportion and number of evading vehicles in each tax class respectively by

$$\bar{u}_h = \frac{1}{n_h} \sum_{k \in s_h} u_k \qquad (18)$$

and

$$\hat{U}_h^3 = N_h \times \bar{u}_h \qquad (19)$$

The corresponding estimates based on the vehicle counts in table 4 are presented in Table 6 below.

Table 6 – "Crude" estimates of evasion ignoring multiple sightings

Tax Class	Crude evasion rate	Registered vehicles (000s)	Crude evasion in stock (000s)
PLG	1,8%	30.928	572
Goods	1,6%	455	7
Motorcycles	14,8%	1.835	271
Buses	0,6%	114	1
Exempt	2,4%	2.081	51
Other	8,8%	121	11
Unknown	10,7%		0
All	2,0%	35.534	704

Source: estimates obtained using estimators (17) – for all vehicles – and (18) for vehicles in each tax class, applied to vehicle counts provided in table 4 and numbers of registered vehicles from table 5 in (Department for Transport, 2007).

The sample means $\bar{u}_h$ are also described as *crude measures of evasion* in the publication presenting the results of the VED evasion – see tables 11 to 13 in appendix B of (Department for Transport, 2007). These sample means estimate the corresponding population means defined as $\overline{U}_h = U_h / N_h$, for h=1,…,6.

Note that each vehicle counts only once in determining N_h and n_h despite the fact that some vehicles may be sighted more than once in the RTOS. The population counts N_h are obtained from the VLA database and hence are known without sampling error. Note also that the sample mean of the u_k is the same as the sample mean of the x_k, i.e. $\bar{u}_h = \bar{x}_h$, because only vehicles which are in use are sampled (y_k=1 for all $k \in s$). This holds both in each tax class and overall.

The poststratified estimator obtained by adding up the estimates of evading stock in each tax class is given by

$$\hat{U}^4 = \sum_{h=1}^{6} \hat{U}_h^3 = \sum_{h=1}^{6} N_h \times \bar{u}_h \qquad (20)$$

and would evaluate to 912 thousand vehicles. The corresponding evasion rate would be obtained as $\hat{U}^4 / N = \left(\sum_{h=1}^{6} N_h \times \bar{u}_h\right) / N = 912 / 35{,}534 = 2.6\%$.

Note the large reduction in the estimates for the evasion in the Motorcycles presented in table 6 in comparison to the corresponding estimates in table 5. For most other tax classes (except Unknown) the direction of change was the opposite, namely, estimates obtained ignoring the repeated sightings (table 6) were generally slightly bigger than those obtained considering the repeated sightings (table 5).

Of all the estimators considered so far, under the working assumption H_2), statistical theory would recommend using (18) and (19) for the evasion by tax class, and (20) for the overall evasion. These are direct estimators of the relevant target parameters (*evasion in stock*), and in the absence of additional information, would be the more precise estimators available. Note that the estimates presented as crude measures of overall evasion by tax class (PLG, Goods, and Motorcycles) in table 11 in Appendix B of (Department for Transport, 2007) were based on the poststratum means (18), but the crude overall evasion reported in §2.2 is based on (11) and not on the poststratified estimator based in (15).

Hence these so-called “crude measures” of evasion could be improved by adopting a coherent set of estimators, all of which are easily calculable given the same survey data, but which would provide more precise estimators than the ones currently employed for the overall evasion.

However, the assumption of vehicle selection / observation with equal probabilities is not easily justifiable given the survey circumstances. Under §2.4 of (Department for Transport, 2007) the idea that sightings resulting from different survey observation sites need to be weighted is introduced. The weighting, carried out separately by region and road class, is described as needed "to reflect the average daily traffic levels of those sites to give an estimate of evasion by road class. National estimates are then made by taking into account different levels of vehicle mileage on different types of road for each region. All these weightings are also split down according to the main vehicle tax classes". The weighted estimates are presented in the publication as measures of "evasion in traffic" – see section 3 and tables 1 till 4 of (Department for Transport, 2007).

Hence the estimation approach was modified to account for the fact that different observation sites have different volumes of traffic. Although the volume of traffic is the key weighting variable utilized to modify the crude estimators, no information is available regarding how the survey observation sites were selected, hence it was not possible to obtain *selection probabilities* for the observation sites. As a consequence, we cannot compare the weights currently used with weights that might be derived from inclusion probabilities for the different observation sites.

A last remark is that these so-called crude estimates of evasion in stock were not used to produce corresponding estimates for VED evasion (*evasion in tax*) in the published report for 2006 – see (Department for Transport, 2007).

3.3 Estimation of evasion in traffic and in stocks

A large portion of the estimates published on the VED evasion report refer to the concept of *evasion in traffic* – see section 3 of (Department for Transport, 2007). In order to define the relevant target parameters corresponding to *evasion in traffic* the framework introduced in section 2 needs to be extended by defining one additional variable. Let q_k denote the mileage done by vehicle k during the whole of the survey observation period. Then the *total mileage for registered vehicles in use* during the survey period is given by

$$Q = \sum_{k \in P} q_k \, y_k \qquad (21)$$

and the corresponding total mileage for vehicles in the various tax classes are given by

$$Q_h = \sum_{k \in P_h} q_k \, y_k \quad \text{for } h = 1,\ldots,6 \,. \tag{22}$$

Table 7 below contains some additional population parameters which can be defined with respect to this new variable, recognizing the domains defined by separating evading and non-evading vehicles. Using these parameter definitions, the *evasion in traffic* is now defined as the ratio of total mileage run by evading vehicles divided by the total mileage run by all vehicles in use, namely:

$$E = Q^U / \left(Q^U + Q^L \right) \tag{23}$$

Table 7 – Some additional parameter definitions for mileage considering domains

Parameter	Description
$\overline{Q} = \sum_{k \in P} q_k \, y_k / \sum_{k \in P} y_k$	Average mileage run by registered vehicles in use
$\overline{Q}_h = \sum_{k \in P_h} q_k \, y_k / \sum_{k \in P_h} y_k$	Average mileage run by registered vehicles in use by tax class
$Q^U = \sum_{k \in P} q_k \, u_k$	Total mileage run by evading vehicles
$Q_h^U = \sum_{k \in P_h} q_k \, u_k$	Total mileage run by evading vehicles in tax class *h*
$Q^L = \sum_{k \in P} q_k \, (1 - u_k)$	Total mileage run by non-evading vehicles
$Q_h^L = \sum_{k \in P_h} q_k \, (1 - u_k)$	Total mileage run by non-evading vehicles in tax class *h*
$\overline{Q}^U = \sum_{k \in P} q_k \, u_k / \sum_{k \in P} u_k = Q^U / U$	Average mileage run by evading vehicles
$\overline{Q}_h^U = \sum_{k \in P_h} q_k \, u_k / \sum_{k \in P_h} u_k = Q_h^U / U_h$	Average mileage run by evading vehicles in tax class *h*
$\overline{Q}^L = \sum_{k \in P} q_k \, (1 - u_k) / \sum_{k \in P} (1 - u_k) = Q^L / (N - U)$	Average mileage run by non-evading vehicles
$\overline{Q}_h^L = \sum_{k \in P_h} q_k \, (1 - u_k) / \sum_{k \in P_h} (1 - u_k) = Q_h^L / (N_h - U_h)$	Average mileage run by non-evading vehicles in tax class *h*

Using the parameter definitions in table 7, this ratio can also be written as:

$$E = \overline{Q}^{U} \times U / \left[\overline{Q}^{U} \times U + \overline{Q}^{L} \times \left(N - U \right) \right] \tag{24}$$

which corresponds to the population parameter that the current methodology defines as "*proportion of evading traffic* (q)" under §D5 in appendix D of (Department for Transport, 2007).

Denoting by

$$F = \overline{Q}^{U} / \overline{Q}^{L} \tag{25}$$

the *relative (average) mileage of evading to non-evading vehicles in use*, and dividing both the numerator and the denominator of (25) by $\overline{Q}^{L}$, it follows that:

$$E = F \times U / \left[F \times U + \left(N - U \right) \right] \tag{26}$$

Recalling that $\overline{U} = U / N$ is the proportion of evading vehicles, and dividing numerator and denominator on the right-hand side of (26) by *N* leads to

$$E = F \times \overline{U} / \left[F \times \overline{U} + \left(1 - \overline{U} \right) \right] \tag{27}$$

Solving (27) for $\overline{U}$ results in

$$\overline{U} = E / \left[E + F \left(1 - E \right) \right] \tag{28}$$

This expression connects the proportion of evading vehicles $\overline{U}$ (or the *evasion in stock*) to the proportion of *evading traffic E* which the survey 'observes', according to (Department for Transport, 2007). Expression (28) corresponds, in our notation and as a population parameter, to the expression for "p" in paragraph D5 of appendix D in (Department for Transport, 2007).

Similar derivations could be used to express the evasion in stock in each tax class $\overline{U}_h = U_h / N_h$ as a function of evasion in traffic per tax class $E_h = Q_h^U / \left(Q_h^U + Q_h^L \right)$ and relative mileage by tax class $F_h = \overline{Q}_h^U / \overline{Q}_h^L$, leading to

$$\overline{U}_h = E_h / \left[E_h + F_h \left(1 - E_h \right) \right] \quad \text{for } h=1,\ldots,6 \tag{29}$$

Hence to estimate the target *evasion in stock* $\overline{U}$ (or alternatively, U, since N is known), the survey needs to estimate two quantities: E (the *evasion in traffic*) and F (the *relative mileage* of evading to non-evading vehicles in use). Due to the nature of the survey observation process, however, the mileage run by each vehicle is not observed. Note that both parameters E and F depend on the unobserved mileage of vehicles in use. The only indirect evidence regarding mileage done by each vehicle provided from the survey is obtained from the repeat sightings of the same vehicle.

Quoting (Department for Transport, 2007), p. 18:

"D1 It is plausible that evading vehicles may not average the same mileage as properly licensed vehicles. A method has therefore been developed to test and model this effect. Note that this model does not assume that relative average mileage is different but rather tests whether this is the case and estimates the difference."

"D2 The model is based on the assumption that vehicles that travel further on average will be observed more frequently in the survey. Although repeat sightings are relatively rare, the survey is large enough to ensure that enough repeat sightings are observed to give good results for the main vehicle classes. For less common tax classes, the relative mileage is estimated in the same way but is subject to larger statistical errors."

This is then the third working assumption required by the current survey estimation procedure – see §1 in Appendix C of (Department of Transport, 1984), and more recently, §D2 in (Department for Transport, 2007):

> H_3) The average number of sightings of a given vehicle is proportional to its mileage run during the survey period.

In terms of population parameters, the above hypothesis implies that:

$$\overline{R} = \alpha \times \overline{Q} \tag{30}$$

$$\overline{R}^U = \alpha \times \overline{Q}^U \tag{31}$$

$$\overline{R}^L = \alpha \times \overline{Q}^L \tag{32}$$

where

$$\overline{R} = \sum_{k \in P} r_k \, y_k \Big/ \sum_{k \in P} y_k \tag{33}$$

$$\overline{R}^U = \sum_{k\in P} r_k\, u_k \Big/ \sum_{k\in P} u_k \tag{34}$$

$$\overline{R}^L = \sum_{k\in P} r_k\,(1-u_k) \Big/ \sum_{k\in P} (1-u_k) \tag{35}$$

are the averages of *numbers of sightings* for all vehicles, evading vehicles, and non-evading vehicles respectively.

Consequently, under H_3) it follows that the relative mileage of evading traffic can be written as:

$$F = \overline{Q}^U / \overline{Q}^L = \frac{\overline{R}^U \big/ \alpha}{\overline{R}^L \big/ \alpha} = \overline{R}^U \Big/ \overline{R}^L \tag{36}$$

Similar expressions would apply for the corresponding relative mileages by tax class, namely:

$$F_h = \overline{Q}_h^U / \overline{Q}_h^L = \overline{R}_h^U / \overline{R}_h^L \tag{37}$$

where

$$\overline{R}_h = \sum_{k\in P_h} r_k\, y_k \Big/ \sum_{k\in P_h} y_k \tag{38}$$

$$\overline{R}_h^U = \sum_{k\in P_h} r_k\, u_k \Big/ \sum_{k\in P_h} u_k \tag{39}$$

$$\overline{R}_h^L = \sum_{k\in P_h} r_k\,(1-u_k) \Big/ \sum_{k\in P_h} (1-u_k) \tag{40}$$

Given these alternative expressions for F and F_h the survey data can be used to obtain corresponding estimates based on observed numbers of *vehicle sightings*. Basically, the estimators are simply the sample averages of the numbers of sightings of evading and non-evading vehicles in each tax class, defined as:

$$\bar{r}_h^U = \sum_{k\in s_h} r_k\, u_k \Big/ \sum_{k\in s_h} u_k \tag{41}$$

$$\bar{r}_h^L = \sum_{k\in s_h} r_k\,(1-u_k) \Big/ \sum_{k\in s_h} (1-u_k) \tag{42}$$

leading to the following estimator for F_h:

$$\hat{F}_h = \bar{r}_h^U / \bar{r}_h^L \tag{43}$$

In fact, the estimator currently used for F_h is (43) times a bias correction term, which is used because the ratio of sample averages is not unbiased for the corresponding ratio of population averages. This bias correction is ignored in this report, because the adjustment is negligible given the very large sample sizes of the RTOS.

Note that the use of unweighted sample averages of the numbers of sightings per vehicle in (43) implies treating the sample of vehicles as if selected with equal probabilities, as was the case for the crude estimates of evasion (section 3.2).

To complete the estimation of $\overline{U}_h = E_h / \left[E_h + F_h (1 - E_h) \right]$, estimates of the evasion in traffic in each tax class E_h are needed. Estimates $\hat{E}_h$ are obtained using the vehicle sightings captured in the RTOS, and weighting them with information on traffic volume derived from the National Road Traffic Survey (Department of Transport, 2006). The weighting scheme used to obtain these estimates is quite complex, and its understanding required examining the actual spreadsheets used for calculation of the *evasion in traffic* estimates, since proper documentation is not available. A detailed description of the current weighting is being prepared, and will be the topic of a second report as part of this review.

The estimator currently used for *evasion in stock* is then defined as:

$$\hat{U}_h^5 = N_h \times \hat{\overline{U}}_h = N_h \times \hat{E}_h / \left[\hat{E}_h + \hat{F}_h \left(1 - \hat{E}_h \right) \right] \qquad (44)$$

where $\hat{E}_h$ is the weighted estimate of the *evasion in traffic* in tax class *h*.

The values of the corresponding totals and rates of evasion in stocks are provided as the main estimates in section 4 (tables 5, 6a and 6b) of (Department for Transport, 2007). These are then used as the multipliers for the average tax value per tax class in (9) to obtain the estimates of VED evasion by tax class provided in table 7 of (Department for Transport, 2007). The average tax values per vehicle tax class are computed from the records in the VLA database, but are not described in detail in the available documentation. Again this is a part of the estimation process that needs to be addressed in the second report which is part of this review.

3.4 Estimation of evasion in traffic and in stocks: the Negative Binomial model

The justification for the estimation approach currently utilised also follows from a model for the number of sightings R_k of a vehicle selected at random from the population of vehicles. The first model to be considered was a Poisson model, briefly described in page 21

of (Department of Transport, 1995). This model can be derived from the following assumptions:

A1) The average mileage of vehicles during the survey period varies with whether the vehicle is evading or not, and is $\overline{Q}^U$ for evading vehicles, and $\overline{Q}^L$ for non-evading vehicles;

A2) The probability of sighting a vehicle travelling unit distance (one mile) is A;

A3) The probability of more than one sighting of a vehicle travelling unit distance (one mile) is negligible compared to A;

A4) The numbers of sightings in non-overlapping miles are independent.

Under these assumptions, Theorem 7 of (Mood, Graybill and Boes, 1974) states that the number of sightings R_k for a randomly selected vehicle k has Poisson distribution with parameter given by A times the average mileage of the corresponding group. Denoting by $P(\lambda)$ the Poisson distribution with parameter $\lambda > 0$, its probability density function is given by:

$$P[R = r] = \frac{\lambda^r \, exp(-\lambda)}{r!} \quad \text{for r=0,1,…} \tag{45}$$

The mean and variance of the $P(\lambda)$ distribution are both equal to λ. Because the distribution of the observed number of sightings starts with 1 (the vehicle must be sighted at least once), the Poisson model is actually fitted to the *number of repeat sightings* of the vehicle, equal to the number of sightings minus 1. Then the model can be summarised as follows:

$$[R_k - 1 | u_k = 1] \sim P(\lambda^U) \text{ with } \lambda^U = A\overline{Q}^U - 1 \text{ if vehicle } k \text{ is evading} \tag{46}$$

$$[R_k - 1 | u_k = 0] \sim P(\lambda^L) \text{ with } \lambda^L = A\overline{Q}^L - 1 \text{ if vehicle } k \text{ is not evading} \tag{47}$$

Assumption A1 specifies that the distribution of mileage of vehicles has different means depending on whether the vehicles are evading or not. As a consequence, the distributions of the numbers of sightings will also have different means for evading and non-evading vehicles, a fact for which there is supporting evidence in the various previous editions of the RTOS. This can also be tested / verified using current survey data.

Assumption A2 may not be adequate, because different vehicles may be kept at different distances from the selected observation sites. In fact, survey observations of the distributions of the numbers of repeat sightings per vehicle have shown *over-dispersion* with respect to the Poisson distribution derived from assumptions A1-A4. This is easily verified by noting that the sample estimates of variance of the numbers of sightings per vehicle are larger than sample estimates of the corresponding estimates of the mean, which is not in agreement with the Poisson model, for which the mean and variance are the same. See sample mean and variance for the distributions of the number of repeat sightings in Table 8 below.

Table 8 – Sample mean and variance for numbers of repeat sightings by tax class

Tax Class	Mean		Variance	
	Licensed	Evading	Licensed	Evading
PLG	0,1512	0,0704	0,1991	0,0949
Goods	0,1966	0,0778	0,2566	0,1030
Motorcycles	0,0851	0,0266	0,0982	0,0273
Buses	0,4366	0,0833	1,1622	0,0786
Exempt	0,1660	0,0818	0,2486	0,1650
Other	0,1503	0,0286	0,1941	0,0280

Source: RTOS2006 model fitting spreadsheets provided by DfT.

To deal with this perceived difficulty, an alternative model was proposed for the distribution of the number of vehicle repeat sightings. The alternative model can be derived from the alternative sets of assumptions:

B1) $R_k - 1 \left| \lambda^U \sim P\left(\lambda^U\right) \right.$ if vehicle k is evading; $R_k - 1 \left| \lambda^L \sim P\left(\lambda^L\right) \right.$ if vehicle k is not evading; that is, the conditional distributions of the numbers of repeat sightings per vehicle given the average rate of repeat sightings in each vehicle group are Poisson, with the group's average rate as parameter;

B2) $\lambda^U \sim GAMMA(c_U ; \mu_U)$ and $\lambda^L \sim GAMMA(c_L ; \mu_L)$, namely the average rates of repeat sightings are distributed as GAMMA with parameters varying by group (evading and non-evading vehicles).

Here *GAMMA*(c ; μ) denotes the GAMMA distribution with parameters c and μ, with probability density function given by:

$$f(\lambda|c;\mu)=\left(\frac{c}{\mu}\right)^{c}\frac{\lambda^{c-1}}{\Gamma(c)}\exp\left(-\lambda\frac{c}{\mu}\right) \text{ for } l>0,\ c>0 \text{ and } \mu>0. \qquad (48)$$

The mean and variance of the *GAMMA*(c ; μ) distribution are μ and μ^2/c respectively.

Under the alternative assumptions B1 and B2, it follows that the unconditional distribution of the number of repeat sightings of a vehicle is Negative Binomial – see section 1.6 of (Zelterman, 2004). The probability density function of the Negative Binomial distribution with parameters c and μ – denoted *NB*(c ; μ) – is given by:

$$Pr(R-1=r \mid c;\mu)=\frac{\Gamma(c+r)}{r!\Gamma(c)}\left(\frac{c}{\mu+c}\right)^{c}\left(\frac{\mu}{\mu+c}\right)^{r} \text{ for } r=0,\ 1,\ \ldots \qquad (49)$$

The *NB*(c ; μ) distribution has mean and variance given by μ and $\mu\,[1+(\mu/c)]$ respectively. Note that the variance is larger than the mean, which is why this distribution is often used to model count data with over-dispersion.

The observed data on the distributions of number of repeat sightings per vehicle in the RTOS can then be used to estimate the model parameters for both evading $(c_U;\mu_U)$ and non-evading $(c_L;\mu_L)$ vehicles.

Under the original Poisson model, the parameters satisfy the following relationship:

$$\frac{\mathrm{E_M}[R_k \mid u_k=1]}{\mathrm{E_M}[R_k \mid u_k=0]}=\frac{1+\lambda^U}{1+\lambda^L}=\frac{1+A\overline{Q}^U-1}{1+A\overline{Q}^L-1}=\frac{\overline{Q}^U}{\overline{Q}^L} \qquad (50)$$

where $\mathrm{E_M}$ denotes expectations taken with respect to the model distributions. A similar relationship holds under the Negative Binomial model, because the mean of the unconditional distribution of the number of repeat sightings is the same as the mean of its conditional distribution given the rate:

$$\frac{\mathrm{E_M}\left[R_k \mid u_k = 1\right]}{\mathrm{E_M}\left[R_k \mid u_k = 0\right]} = \frac{1+\mathrm{E_M}\left[\lambda^U\right]}{1+\mathrm{E_M}\left[\lambda^L\right]} = \frac{1+\mu_U}{1+\mu_L} = \frac{\overline{Q}^U}{\overline{Q}^L} \tag{51}$$

because $1+\mu_U = A\overline{Q}^U$ and $1+\mu_L = A\overline{Q}^L$.

Now the population values of the parameters on the numerator and denominator of the ratio on the left hand side of (50) or (51) are $\overline{R}^U$ and $\overline{R}^L$ respectively, and the more sophisticated model assumptions provide a second justification for the assumed relationship $F = \overline{Q}^U / \overline{Q}^L = \overline{R}^U / \overline{R}^L$. Analogous derivations produce the equivalence of the average mileage ratios to the average number of sightings ratios for each tax class.

This implies that the sample observations on the distributions of the numbers of vehicle sightings (which provide estimates for $\overline{R}^U$ and $\overline{R}^L$) can be used to estimate the ratio of average mileage run by evading and non-evading vehicles ($\overline{Q}^U / \overline{Q}^L$) which cannot be estimated directly, since the vehicle mileages are unobserved.

The methodology adopted currently for the evasion in stocks estimation fits Negative Binomial models to the distributions of the numbers of repeat sightings separately for evading and non-evading vehicles, within each tax class. The model fitting is carried out using a simple *Method of Moments* approach, applied separately for vehicles within each tax class *h*. Denoting by $\hat{S}^2_{Uh}$ the sample variance of the numbers of repeat sightings of evading vehicles in tax class *h*, the method of moments fits the *NB*(c ; μ) distribution to vehicles of this group by solving equations like

$$\tilde{\mu}_{Uh} = \bar{r}^U_h - 1 \tag{51}$$

$$\tilde{\mu}_{Uh}(1+\tilde{\mu}_{Uh} / \tilde{c}_{Uh}) = \hat{S}^2_{Uh} \tag{52}$$

where $\tilde{c}_{Uh}$ and $\tilde{\mu}_{Uh}$ are the method of moments estimators of c and μ for evading vehicles in tax class *h*.

Solving this system of equations for $\tilde{c}_{Uh}$ yields

$$\tilde{c}_{Uh} = \left(\bar{r}^U_h - 1\right)^2 \Big/ \left[\hat{S}^2_{Uh} - \bar{r}^U_h + 1\right] \tag{53}$$

Note that the method of moments estimator for c_{Uh} is only feasible if the sample variance is larger than the sample mean, because otherwise it will take a negative value. Hence a simple diagnostic for the model fit must be performed before proceeding. If the sample variance is not larger than the sample mean, fit the Poisson model. Otherwise, fit the Negative Binomial model.

More elaborate model checking may be performed by computing goodness of fit statistics under both models, but this simple rule is required to avoid trying to fit the Negative Binomial when this model would not be supported by the sample data at hand.

Similar procedures lead to the parameter estimators for the group of licensed vehicles. Note that the outcome of this estimation process leads to the same estimator for F_h as already presented in (43).

4 Critical review of the current estimation approach

This section contains a summary of the findings resulting from the review of the methodology adopted to estimate vehicle evasion in stock and VED evasion. The key question initially asked referred to the use of the Negative Binomial distribution to model repeated vehicle sightings. This model was developed on the basis of assumptions which are strong, but which appear justifiable in the present survey scenario. The model does follow through if the assumptions are made, and there is no error there.

The model has been fitted to survey data for several subgroups: licensed or unlicensed vehicles grouped by tax class. For some of these subgroups, the sample sizes are insufficient to warrant safe inference (in particular, this is the case for Unlicensed vehicles in the Buses and Other tax class categories) – see the highlighted cells in table 4. Now the model fitting is only relevant if licensed and unlicensed vehicles within the same tax class can be fitted separately, because this is what is required to obtain the relative mileage ratio estimates (43). The very small sample sizes for the unlicensed vehicles in the Buses and Other tax classes implies that these two tax classes have to be combined with some other tax class before the model is fitted. But this is contradictory with the idea that justifies the form of the estimator for VED evasion in (9), which relies on the assumption that values of tax for vehicles in each tax class are approximately constant.

The models are presently fitted using the Method of Moments estimators for the two relevant parameters of the Negative Binomial distribution. This can be improved upon by using the Maximum Likelihood method to fit the models.

However, the estimates for the mean parameter under both methods are identical for practical purposes, and since the mean is the only parameter required to perform the adjustment that converts evasion in traffic into evasion in stocks, the method used to fit the model does not impact on the final estimates of evasion in stocks.

The Negative Binomial model was fitted to all subgroups, but for some of them, this model is clearly inadequate. As mentioned in section 3.4, this happens when the sample variance is smaller than the sample mean. This problem was observed for some model adjustment cells (groups of vehicles in the same tax class and licensing status) where the method of moments' estimator for the *c* parameter would be negative, namely evading (unlicensed) vehicles in the Buses and Other tax classes. Not a surprise that the model fit problems appeared exactly for the same adjustment cells with very small samples.

The model fitting procedures adopted by the DfT team must ensure that the model fitting outcome is examined to detect such cases, which require that a different model is fitted to the data. In both cases where the Negative Binomial was not adequate, the Poison model could provide a simple alternative. Although this problem does not affect the point estimates of evasion in stocks, they illustrate the large uncertainty that estimates in cells with such small samples will have. Dependence on these unreliable estimates is an undesirable feature of the current approach.

The observed distributions of the number of repeat sightings for vehicles in the different tax classes were quite skewed for some vehicle subgroups. A small potential model fit improvement may be made by limiting the maximum number of repeat sightings considered (for example, for PLG there are cases of up to 17 repeated sightings of a single vehicle, but for model fitting purposes, one could consider that counts of 10 and above are all grouped into a single class). This corresponds to some form of winsorization of the data, and would provide some robustness to the model fitting, although at the expense of some bias. A

Maximum Likelihood approach that fits the models using this truncation was developed and used for some analysis, but again the differences observed were not large, if the models are fitted only for subgroups with sufficiently large sample sizes.

Chi-square statistics computed to check the goodness of fit were not always indicative of good model fit for some of the subgroups where the model was fitted. However, the estimates for evasion in stock are not overly dependent on the form of the distribution used to fit the distribution of repeated vehicle sightings. The crucial assumption on which the adjustment that converts evasion in traffic into evasion in stocks relies is that the average number of sightings of a given vehicle is proportional to its mileage (see H_3). This hypothesis is untestable from the survey data because the vehicle mileage is not observed. However, there might be information in the VLA database which might be used for this purpose. In fact, the first time that this working assumption was adopted – see §4 in Appendix C of (Department of Transport, 1984) – some direct survey evidence on the adequacy of this hypothesis was obtained by a postal survey of keepers of heavy goods vehicles. This is an area which will be further investigated in part 2 of the review.

There is one key point of concern not addressed so far. The estimation of the *evasion in traffic* (namely, the E_h) takes account of the weighting of traffic in roads of different types in order to reflect different volumes of traffic. The estimation of the adjustment factors F_h used to convert evasion in traffic into evasion in stocks ignores the weighting altogether and treats the sample of vehicles as a SRSWR. The combination of a traffic weighted estimator $\hat{E}_h$ with an unweighted estimator $\hat{F}_h$ to obtain the evasion in stock $\hat{U}_h^5$ using (44) does not follow from statistical estimation theory in a natural way. This problem needs to be addressed, so that the procedure used to estimate evasion in stock uses the available data in a coherent way: either the weighting is essential, and should be considered for both components, or if it is not essential, perhaps there are simpler yet efficient approaches to estimate the evasion in stock and in tax.

Another issue is that the approach based on first estimating evasion in traffic and then use the ratio of average mileage of evading to non-evading vehicles to 'convert' this estimate

into an estimate of evasion in stock is quite complex. It transforms the problem of estimation of a single quantity (namely, U_h) which could be based entirely into information directly observed in the RTOS surveys, into the estimation of two quantities (E_h and F_h) which are defined on the basis of a variable which is not observed as part of the survey process. Hence the dependence of the latter on the strong, yet unverifiable assumption H_3).

The poststratified estimators presented in section 3.3 would provide a much simpler alternative, which also deals with the issue of multiple sightings by using a single (non-repeat) observation for each vehicle. Further investigation is required to assess whether these estimators are indeed preferable to those currently used for evasion in stock. However, it can already be suggested that these should at least replace the ones currently used as 'crude measures' of evasion.

Documentation about the estimation procedures is scarce and this is one of the areas to be targeted for improvement in future editions of this survey, if not for external publication, at least for the benefit of staff working on the survey. Details of the motivation behind the various alternative estimators were provided here to contribute towards this goal.

These are the initial findings, and do not represent a complete and definitive analysis of the methodology. They are provided as a point of reference from which decisions about how to continue the review can be made. Further investigation is needed about:

1) How weights are obtained and used to estimate evasion in traffic;
2) Whether there are alternatives to the general form of the estimator (9) used to estimate VED evasion, which are less dependent on the working assumption H_1);
3) Whether the current sampling design of the Roadside Traffic Observation Survey needs revision or improvement, and how it generates sampling probabilities for the different sites selected for survey observation;
4) The potential impact of different ways of aggregating the vehicles in subgroups prior to the model fitting and calculation of the adjustment factors to convert evasion in traffic into evasion in stocks;
5) The estimation of precision under the various estimation approaches.

5 References

Boucher, A., and Hird, D. (2007). VED evasion estimates - methodology review. D. f. Transport (ed), 5: Department for Transport.
Department for Transport (2007). Transport Statistics Bulletin: Vehicle Excise Duty Evasion 2006. D. f. Transport (ed), 23: Department for Transport.
Department of Transport (1984). Vehicle excise duty evasion in Great Britain 1984/85. In *Statistics Bulletin*, D. o. Transport (ed), 39. London: Department of Transport.
Department of Transport (1995). Vehicle excise duty evasion in Great Britain: 1994/95. In *Transport Statistics Report*, D. o. Transport (ed), 29. London: Department of Transport.
Department of Transport (2006). Transport statistics bulletin: Road Traffic Statistics - 2005. D. o. Transport (ed), 32: Department of Transport.
Mood, A. M., Graybill, F. A., and Boes, D. C. (1974). *Introduction to the theory of statistics, third edition*, 3rd edition. Singapore: McGraw-Hill Book Company.
Särndal, C. E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
Zelterman, D. (2004). *Discrete distributions: applications in the health sciences*. Chichester: John Wiley & Sons.