PREDICTING NATIONAL DATA ON THE USE OF PRIVATE VEHICLES

IN CANADA FOR THE 1980-1996 PERIOD:

An Application of the Bayesian Approach of Gibbs Sampling with Data Augmentation

by

NATHALIE BOUCHER

A thesis submitted to the Department of Economics

In conformity with the requirements for

the degree of Doctor of Philosophy

Queen's University

Kingston, Ontario, Canada

January 2003

0-612-80983-8

Abstract

An extended statistical software for the estimation, prediction, and inference of a wide variety of standard econometric models is developed to analyze datasets involving a large proportion of missing information. This relies on Bayesian sampling-based approaches with data augmentation. A generalization of the Bayesian treatment of vector autoregressive models is also considered. As a direct by-product, the proposed methodology is shown to be a natural and effective way to address the problem of data interpolation from intermittent longitudinal surveys which is both conceptually simple, and computationally tractable.

We apply the interpolation methodology to bridge the gap between two national surveys on the use of private vehicles which are six years apart. This allows us to produce quarterly predictions of the three energy components (the average number of vehicles, the average distance travelled by each vehicle, and their weighted fuel consumption rate) for the intermediary period, between the surveys. Separate estimates and predictions are obtained by vehicle type: for cars and for light trucks and vans. The same technique could also be directly implemented in other contexts such as international database comparisons, population censuses, longitudinal labour force surveys, etc.

First of all, survey-based estimates are adjusted with the aim of improving their compatibility. Predicted values for the intermediate period are obtained by means of the Bayesian method of Gibbs sampling with data augmentation. In order to improve efficiency, by making use of all available empirical information, the econometric model is estimated on the basis of data from both surveys, while taking into account the middle period, between the two surveys' sampling periods, for which no data exist.

Based on explanatory variables from external sources, the aggregate simultaneous equations model is formulated to account for the relationships among the energy components. The model takes into account the dynamics involved in the three dependent variable time series. Since the data are aggregated on a quarterly basis, it also captures seasonal variations, in addition to the general trends in the series.

Several alternative specifications are compared to determine the best prediction model. The generalized vector autoregressive model is shown to yield the most precise and reliable results. Convergence of the iterative estimation process and its dependence on prior choices are assessed by means of sensitivity analyses. Complete time series produced by this empirical analysis will provide more accurate data on which the policy makers can rely.

Given that a similar survey is to be done soon, the necessity of obtaining, from such intermittent sources, complete time series estimates for the key variables from a transportation researcher's point of view becomes crucial. However, the proposed interpolation methodology is not a substitute to a well-designed data collection process, but rather a general solution to existing data gaps.

Acknowledgements

I am more than grateful to my thesis co-supervisors, Professor Denis Bolduc at Laval University's Economic Department and Professor Allan Gregory at Queen's Economic Department, for their help in writing my doctoral thesis. Professor Bolduc's council and tutoring were appreciated throughout the research process. Professor Gregory's suggestions allowed me to improve the presentation of the thesis.

The author wishes to thank Professor Lisa Powell from the School of Policy Studies at Queen's University, who acted as an advisor, for providing valuable comments.

Financial support of Quebec Government's *Fonds pour la formation de chercheurs et l'aide à la recherche* (FCAR) during the first four years of my Ph.D. studies is gratefully acknowledged.

I am indebted to the Automobile Mobility Data Compendium (AMDC) at Laval University for providing the technical facilities for the econometric treatment. Special thanks are dedicated to Mrs. Sylvie Bonin, chief analyst at the AMDC, for many useful suggestions.

This work was partly sponsored in part by Natural Resources Canada (NRCan) by the intermediary of the AMDC where I worked as an analyst during the past three years to recently become its Executive Director. I am grateful to NRCan for generously providing me with an almost exclusive access to the data used in the empirical study.

Earlier versions of the study have been presented at annual meetings of the AMDC with its sponsors in 1999, 2000, and 2002. We have benefited from comments by participants in these seminars.

The author also wishes to thank her parents and her boyfriend for their love, care and support without the help of whom she would not have achieved her goals.

Table of Contents

Page

List of Tables

Page

List of Figures

Page

Page

Chapter 1

Introduction

The 1992 Rio meeting witnessed world-wide concern about pollution and its effects on the environment. More concrete engagements have resulted from the Kyoto meeting of the signatories of the United Nations on climate changes in 1997. At that meeting, the Canadian government agreed to reduce its greenhouse gas (GHG) emissions by 6 percentage points below their 1990 level during the 2008-2012 period. In practice, according to Natural Resources Canada (NRCan, 1998), this means that emissions of carbon dioxide (CO_2), the main GHG,[1] must be reduced by nearly 20 percent.

Transportation currently generates more GHG emissions than any other sector --- 27 percent of total Canadian emissions --- almost as much as the residential, commercial and industrial sectors combined.[2] Cars, light trucks and vans alone are responsible for almost 60 percent of total transport GHG emissions. Among those vehicles, private transport accounts for almost 90 percent of all passenger activity, measured by the distance travelled by all passengers. Thus, the private transport sector represents a large share of total GHG emissions.

Consumers' holding and use of motor vehicles have been affected by governmental policies and private sector actions (Ben-Akiva, Manski and Sherman, 1981). Major policy changes that have recently been instituted, or are now being considered, include vehicle fuel economy and emission standards, fuel, vehicle purchase and registration taxes, and safety-related design restrictions. More sporadic actions, such as the institution of exclusive lanes for high-occupancy vehicles and public transport service improvements, are aimed at reducing private vehicle use in congested urban areas. In order to implement careful policy rules to control pollution, the Canadian

[1] According to the Quebec Minister of Environment (1999), CO_2 emissions represented 80 percent of overall world-wide GHG emissions in 1996.

[2] See the Transportation Table's Foundation Paper on Climate Change (1998) for a complete summary of the available information on transport GHG emissions.

government must rely on accurate measures of the actual fuel consumption and the resulting CO_2 emissions attributed to the private transport sector.

An improved understanding of the relationship between energy efficiency, energy use, and GHG emissions will assist policy-makers in developing more effective solutions to the issues of global climate change, urban smog, and sustainable development. In terms of the private transport sector, these solutions involve reducing automobile vehicle CO_2 emissions by encouraging less intensive use and by improving fuel efficiency.

The CO_2 emissions are measured by applying emission factors developed by Environment Canada (EC, 1997) to fuel consumption. Although EC and NRCan rely on slightly different divisions of end-use energy demand by sector,[3] both use data from NRCan's Transportation Energy Demand Model (TEDM) to measure private vehicle fuel consumption and its main components. It is therefore crucial to obtain reliable measures of private transport aggregates in order to estimate accurately present and future levels of fuel consumption, as well as Canadian CO_2, and hence overall GHG, emissions.

The TEDM factorizes fuel consumption as the product of three main components: the number of vehicles, the average distance they travel, and their fuel consumption rates. Each variable is measured independently, based on information available at the national level. However, the current measures are not based on disaggregate empirical data and suffer from severe drawbacks with regard to the limitations and precision of available information, unrealistic underlying assumptions, and related data manipulations.

Two nation-wide surveys on the use of private vehicles provide important and exclusive empirical information to improve upon the actual measurement of private transport aggregates. The *Fuel Consumption Survey* (FCS) sampled cars from the

[3] Those differences do not concern the private transport sector, however.

fourth quarter of 1979 to the last quarter of 1988, but light trucks and vans for a shorter period of time delimited by the fourth quarters of 1981 and 1987. The *National Private Vehicle Use Survey* (NaPVUS) was initiated in the last quarter of 1994, for a period of 24 consecutive months, to collect information on both types of vehicles. These provide, throughout the years, a detailed profile on the use of private vehicles in Canada, excluding Northern Territories. Thanks to NRCan, from the intermediary of the Automobile Mobility Data Compendium (AMDC), we were the first to have access to these data.

From such disaggregate surveys, quarterly estimates of each variable of interest can be derived separately for cars, and for light trucks and vans. However, there is a gap of almost six years between them, representing nearly 34 percent of the cars sample, and 45 percent of the light trucks and vans sample. Furthermore, the two surveys did not sample exactly the same classes of vehicles and failed to collect some relevant information. Given that a similar survey is to be undertaken soon, there is a striking need to reconcile the data and obtain complete time series.

We propose the Bayesian approach as a natural, effective, and relatively simple way to handle the general problem. More precisely, we apply sampling-based techniques, such as the Gibbs sampling algorithm, to estimate a simultaneous equations model involving the three transport aggregates. Adding a data augmentation stage to such an iterative estimation process provides predictions for the missing values as a direct by-product. Data augmentation fills in the missing information in a completely symmetrical and endogenous way, and produces predictions that are coherent with the hypothesized econometric model. Furthermore, the method proves to yield satisfactory results, even in small sample applications involving a significant share of missing information.

We compare several estimation techniques, model specifications, and prior distributions in order to determine the best prediction model. An extended statistical software package, based on Fortran programming, is designed to analyze the various model formulations which are the most frequently encountered in applied econometrics. We propose a generalization of the Bayesian treatment of vector autoregressive models that allows for the inclusion of stochastic variables, in addition to deterministic and lagged dependent variables, in the model formulation. Asymmetry across the system equations is also admissible in this more general setting. The resulting model is shown to yield the most reliable and accurate predictions.

The Bayesian sampling-based techniques with data augmentation have a wide range of applications in all empirical problems involving missing information. For instance, they can be implemented to impute values to unobserved data and obtain complete datasets in disaggregate cross-sectional surveys involving non-responses to certain questions (e.g. respondent in a household survey who refuses to report salary), and panel data (e.g. attrition due to some subjects dropping out before the end of the survey), in addition to aggregate macroeconomic time series (e.g. incomplete national income figures for some countries). They can also be used to endogeinize latent variables (e.g. consumer's utility) in a qualitative and limited dependent variables modelling framework.

The thesis organization is the following. Chapter 2 provides the background motivation for undertaking the present analysis and describes the main problem in more detail. We present a literature review on the modelling of transport aggregates, as well as on the Bayesian sampling-based approaches to estimation and prediction. A description of the empirical data is given in Chapter 3. We outline the differences in the two surveys' sampling methodology and data collection process, and account for them in order to produce more compatible estimates. Chapter 4 explains how the Bayesian sampling-based techniques can be implemented to estimate the simultaneous equations model and derive predictions for the missing values in its dependent variable series. Chapter 5 provides the basic estimation and prediction results. We first test the Bayesian estimation methods' performance based on simulated data, and then apply them to solve our empirical problem. Chapter 6 compares the results obtained from alternative model specifications and estimation procedures in order to determine the most appropriate candidates for our prediction purpose. We address convergence issues and question the underlying prior assumptions for the selected model. Chapter 7 summarizes the main findings and conclusions of the empirical study.

Chapter 2

Literature Review

This chapter reviews the literature on the main themes involved. First, it outlines the background motivations for undertaking the present study. We have already stressed the importance of assessing precise and accurate data on the use of private vehicles for policy making and air pollution control. Section 2.1 examines in detail how these aggregates are actually measured, highlighting the main drawbacks and imprecision of some underlying data and assumptions. Section 2.2 shows how two nation-wide surveys on private vehicle use provide us with exclusive and critical empirical information that can help improve the measurement of private transport aggregates. We argue that the past research in that direction has yielded misleading or imprecise results either because it ignored specifics to each survey sampling field, or because it could not make use of all the empirical evidence.

Having provided the justification and reviewed the challenge involved in undertaking this study, the main empirical problem is outlined in Section 2.3. Specifically, we apply adjustment techniques to reduce the discrepancies between the two survey-based estimates and propose econometric estimation and prediction methods to fill the gap between them using all available empirical information. A summary of the different ways each aggregate is modelled in the transportation literature is described in Section 2.4. Finally, the Bayesian approach and related sampling-based techniques are briefly revisited in Section 2.5. This review leads to a better understanding of the general problem and motivates the Bayesian estimation and prediction methods employed for this analysis.

2.1. Measurement of fuel consumption in the private transport sector

NRCan's Office of Energy Efficiency Demand Policy and Analysis Division publishes, annually, a document entitled *Energy Efficiency Trends in Canada* (e.g. NRCan, 1997). This document aims at analyzing changes in Canadian aggregate energy demand and how they impact on environmental conditions. The empirical facts reported in this publication rely on annual data produced by the Transportation Energy Demand Model (TEDM), which is also used for prediction purposes and policy analysis. The TEDM partitions energy consumption into five main sectors of activity: industrial, commercial, residential, agricultural, and transportation. In this model, each energy sector is analyzed independently of the others by means of a bottom-up model. This is a pyramidal structure in which each stage corresponds to a more or less disaggregated level of the total energy demand for that sector. Sectors are thus further divided into sub-sectors, and the Canadian energy demand is split by province.

The present document is only concerned with the transportation sector. More precisely, while the TEDM that accounts for this sector's demand is further segmented into two main components, namely the private and commercial transportation sectors, we focus on the private sector. This sector essentially comprises road transportation. Therefore, while the TEDM distinguishes among four vehicle modes, road, aviation, rail, and marine, we are solely interested in road. Although road transportation is composed of four vehicle types, cars, trucks and vans, buses, and motorcycles and mopeds, we concentrate on the first two. The bottom-up structure of the TEDM thus allows one to focus on the fuel consumption of cars and light trucks and vans that are used for private purposes in Canada.

One aim of this thesis is to provide NRCan with more accurate data on private transportation. The scope of our analysis is narrowed to private-use cars, light trucks and vans because we make use of empirical data from two surveys covering only these vehicles. Since we have to improve upon the actual transport aggregates data, a natural starting point is to begin by examining how these are currently measured. In particular, we wish to stress the weakness and imprecision involved in these measurements that our empirical analysis will compensate for.

The TEDM relies on a basic identity which decomposes total fuel consumption, C, into the product of its three main components as follows:

$$C=S\left(\frac{D}{S}\right)\left(\frac{C}{D}\right)=S\,\overline{D}\,E \qquad (1)$$

where S stands for the corresponding vehicle stock, and D represents the total distance travelled by these same vehicles. Hence, $\overline{D}$=D/S symbolises the average distance travelled by each of these vehicles, and E=C/D, their fuel efficiency, or weighted fuel consumption rate, expressed in terms of litres by 100 kilometres. Each of these components, as well as their product, is measured separately. The distance variable, whose measurement accuracy is the weakest, is then computed residually to ensure the equality on both sides of the equation.

Data on each of the components of equation (1) are now available for the 1976-1997 period. These are measured on an annual basis for Canada as a whole. For the private road sector, they are disaggregated by province, vehicle type, size, and model year, as well as fuel type. Some of these segmentations have been undertaken quite recently. For instance, over the past year, the disaggregation of Canada into seven regions has been extended to a ten-province breakdown. Over the same period, new sources of information based on registrations have permitted the distinction between private and commercial vehicles, in addition to a revised classification of the light vehicle stock.

Each variable is derived from different (and sometimes various) sources. Some data come from private consultants, but the majority are drawn from Statistics Canada (StatCan) catalogues. Note that none of these rely on disaggregate survey-based information. Fuel consumption estimates are from StatCan's *Quarterly Report on Energy Supply-Demand in Canada* (see catalogues no. 57-003) and are based on fuel sales as reported by Canadian distributors. Lab-tested fuel consumption rates (FCRs) are estimated by the automobile manufacturers as reported in the *Vehicle Fuel Economy and Emissions System* (VFEES) and the *Vehicle Fuel Economy Information System* (VFEIS)[1].

Two sources, both based on provincial registration files, are used in determining the number of vehicles: The *Canadian Vehicle in Operation Census* (CVIOC) provided by Desrosiers Automotive Consultants covers the 1989 to 1997 period, while the *Truck Information Profile* (TIP) provided by Polk Canada only goes from 1994 to 1997. In order to extrapolate the data to the uncovered period, NRCan makes use of new vehicles sales and survival curves[2] to evaluate, respectively, the inflow and the outflow from the previous vehicle stock. StatCan's monthly reports on *New Motor Vehicle Sales* (catalogues no. 63-007-XIB) yield basic evaluations of new vehicle sales. StatCan's annual reports on *Trucking in Canada* (see catalogues no. 53-222-XPB) provide further information on the total sales of new trucks. Finally, predictions are calibrated to meet the general trend in registrations, as displayed in StatCan's annual reports on *Road Motor Vehicles, Registrations* (see catalogues no. 53-219-XIB).

The profusion of sources renders the data reconciliation difficult. Most of these sources collect provincial information which is aggregated afterwards. This raises the necessity of accounting for inter-provincial discrepancies in regulations and vehicle classifications. Apart from fuel consumption data which takes into account inter-provincial transfers, no consideration is devoted to double-counting that may result from vehicles' inter-provincial migrations. According to StatCan standards, provincial

[1] Since 1994, the name of the VFEES database has been changed to VFEIS as it does not include information on carbon dioxide emissions anymore.

[2] Survival curves come from the National Energy Board (NEB) model. We were unable to obtain details on the NEB model from NRCan, apart from being told that this was an old empirical model relying on evidence from the 1970's. It seems that no precise reference exists on the underlying assumptions of this model which is nevertheless widely used internally.

estimates are not reported whenever confidentiality might be violated, leading to downward bias in total estimates.

The fact that data are available at different levels of aggregation, or for distinct categories of vehicles, raises additional merging difficulties. For instance, lab-tested FCRs which are available on an annual basis for each model of vehicle must be merged to new registrations as reported by each province every year. Then, weighted averages of lab-tested fuel efficiency have to be computed to meet each segment of the vehicle stock considered by NRCan. Finally, lab-tested FCRs are adjusted in order to better reflect the on-road vehicle fuel efficiency.

In all of the above mentioned publications, StatCan uses a misleading classification of on-road vehicles distinguishing "passenger automobiles" from "trucks and truck tractors".[3] Taking the terminology as proposed does not allow one to discriminate light trucks from medium and large ones.[4] According to the definitions provided at the beginning of StatCan's catalogues, however, the "passenger automobiles" class includes cars, light trucks and vans that are used for private purposes, while the "trucks and truck tractors" class refers to commercial-use cars and trucks. In this sense, the classification provides a better partitioning of vehicles by use than by type. But this does not even yield a perfect breakdown of the vehicle stock by use because cabs are included in the private vehicle category.

Under this classification, most minivans will be part of the "passenger automobiles" class, as they are mainly used for private purposes. However, in the TEDM, minivans are assigned to light trucks as their fuel consumption better relates to this category of vehicles. This further biases estimates of the number of vehicles

[3] Refer to Boucher (1998a) for a complete explanation.

[4] This appears to be the current practice at NRCan which bases the partitioning between light trucks and medium and heavy trucks on the disputable assumption that it is the same as in the U.S.

according to their type derived from registrations (in favour of a larger car fleet) and yields misleading fuel consumption estimates. For instance, when lab-tested FCRs are merged with registrations of new vehicles, cars (including minivans) weighted fuel efficiency estimates are lower than they should be under the TEDM classification.

The need to have even finer levels of disaggregation to meet NRCan's desired vehicle stratification, often leads to the use of unrealistic or simplifying assumptions, or to adopt ad hoc rules based on more or less arbitrary numbers. For instance, it is assumed that the average distance travelled by private-use vehicles has remained constant since the 1970's. This is, of course, an unrealistic assumption, as average distance travelled has historically increased.[5] Disaggregation of trucks according to their size relies on the U.S. partitioning of the vehicle fleet which does not necessarily apply to Canada. To split fuel consumption according to vehicle types, fixed personal use shares are applied to the total energy consumed, by means of an ad-hoc rule. Each lab-tested FCR is multiplied by an arbitrary factor of 1.2 in order to better reflect the on-road fuel efficiency of light vehicles.[6]

There are other problems with the manipulations operated on NRCan's data, but it is sufficient it to say that these calculations are based on rather unrealistic, unjustified, too simplistic, or at least questionable hypotheses. Their application may result in inaccurate or imprecise estimates of the transport aggregates. Some even lead to sizeable inconsistencies. For instance, survival curves, used to determine how the existing stock ages and eventually lapses, turned out to be completely flat, for some years, as if the vehicle would last forever. This indicates a contradiction between the predicted stock, as derived from the registration files, and the sales of new vehicles, as reported by StatCan.

[5] This is confirmed by survey-based data used in Chapters 3 through 6.
[6] Apparently, the 1.2 factor also comes from the NEB model.

As argued throughout this section, actual measurements of the private transport aggregates suffer from several serious drawbacks. The following section shows how the release of new empirical information opens up avenues for improvement.

2.2. Evidence from survey-based data

Two national surveys have been undertaken by StatCan in order to collect empirical information on the private transport aggregates. These surveys will help governments and interested parties to better monitor and analyse trends in fuel consumption and GHG emissions from the private vehicle sector. The *Fuel Consumption Survey* (FCS) took place in the 1980's and covers a period of almost ten years. The *National Private Vehicle Use Survey* (NaPVUS) was then undertaken in the mid-1990's for a period of 24 consecutive months.

The FCS and the NaPVUS are primary sources of longitudinal information in studying the vehicle fleet in Canada. An Automobile Mobility Data Compendium (AMDC) report (Boucher, 1998a) showed how few sources of this kind were available and how difficult it was to reconcile data from some of those sources. Therefore the generation of complete time series based on the two national private vehicle use surveys will be very useful. These surveys will be described in detail in Chapter 3. Here, we simply present the first empirical results based on both survey estimates that are relevant for this study.

2.2.1. Natural Resources Canada's news release

Following the first estimation results derived from the NaPVUS for the fourth quarter of 1994, NRCan issued a news release on October 2nd, 1996. The news release provided the first comparisons between the two surveys' estimates. The release contrasted FCS estimates for the fourth quarter of 1987, last year during which the FCS surveyed both cars and light trucks/vans, with the first NaPVUS estimates.

Initial comparisons outlined important increases in private vehicle use in Canada during the six years separating the two survey sampling periods. According to the news release, while fuel efficiency witnessed some noteworthy improvements (of 2.1 percent a year, on average), the average distance travelled by Canadians increased by 17.2% from the last quarter of 1987 to the last quarter of 1994, for an annual increase of 2.5%. The huge gap worried NRCan considerably, considering the fact that intensity rose faster than efficiency over the same period.

2.2.2. First forecasting trials

A previous study undertaken by the AMDC for NRCan illustrates how difficult it is to extrapolate survey-based estimates to the intermediary period, between the two surveys. This study, which was only concerned with predicting the fuel consumption of private-use vehicles (cars and light trucks and vans all together) up to 1994, revealed that predictions solely based on the FCS data were highly unreliable given that the prediction period was almost as long as the historical records. Furthermore, whenever the first results based on the NaPVUS data for the fourth quarter of 1994 became available, predictions were found to underestimate substantially the fuel consumption for that quarter.

First attempts to predict quarterly fuel consumption based on the FCS data only are summarised in Bonin and Bernard (1996). Linear regression models, including lagged dependent variables, and autoregressive moving average (ARMA) models were specified. In addition to lagged dependent variables, other explanatory variables such as net quarterly fuel sales (pump sales), average unleaded fuel price, unemployment rates, and the gross domestic product (GDP) of some industries were included in the specification. The variables that were found to have a significant effect (at conventional levels) on fuel consumption are the four-period lagged fuel sales and the GDPs of the retail trade and service industries.

As will be argued in Chapter 3, the conclusions of the news release and Bonin and Bernard's report are somewhat mitigated by the fact that they ignore the inherent differences between the two surveys' sampling methodology. As will be shown in Chapter 3 once these differences are accounted for, the estimates from both surveys are more easily reconcilable. One important contribution of the present study is in the adjustment methods aimed at reducing such discrepancies in order to produce compatible estimates. The next section outlines the purpose of our empirical analysis and the way it will proceed to achieve its goals.

2.3. Bridging the gap between the two surveys

We propose a way to fill the data gap between the two national surveys on the use of private vehicles. Our goal is to produce quarterly predictions of the three energy components, the average number of vehicles, the average distance travelled by each of these vehicles, and their weighted fuel consumption rate for the period between the FCS and the NaPVUS. From these, consistent and complete estimates also can be obtained for the total distance travelled by all vehicles and their overall fuel consumption. Separate estimates and predictions will be obtained by vehicle type: for cars and for light trucks and vans in isolation. These will provide more accurate data for use by NRCan and others.

In order to improve efficiency by making use of all available empirical information, the econometric model will be estimated on the basis of data from both surveys, accounting for the fact that no data exist for the middle spell, between the two surveys' sampling periods. Since survey data are not available for the intermediary period, complete series of explanatory variables will be drawn from external sources to help forecast. Note that the primary interest here is to obtain accurate predictions of the transport aggregates, but not to estimate a structural econometric model for such variables.

Aggregate demand modelling is considered since only the NaPVUS information would allow for a more disaggregated, household-based approach, which would not be appropriate for prediction purposes, anyway. First of all, survey-based estimates must be modified by a series of adjustments to make them more comparable. The prediction model must take into account the dynamics involved in the three dependent variable time series. Since the data are aggregated on a quarterly basis, it must capture the seasonal variations, in addition to the general trends in the series.

The next two sections explain in more detail how we proceed. We revisit the literature on modelling the transport aggregates in order to determine which econometric model is the most appropriate for our purpose. Then, we review the econometric methods that will be used to estimate the model and derive predictions for the missing data.

2.4. Modelling fuel consumption

Transportation literature has proposed many econometric models for fuel consumption (C) and its three main components: vehicle stock (S), distance travelled (D) and fuel efficiency (E). While the primary interest is on modelling the aggregate time series, we also review some of the related disaggregate studies in order to draw a more general picture of the issues involved in modelling these series and to get some insights on the explanatory variables that such models usually involve.

Dahl and Sterner (1991) provide a survey of empirical work on fuel consumption. An indirect way of modelling fuel consumption (e.g. see Schimek, 1996, for an aggregate application and Hensher, 1986, for a disaggregate one) is by modelling simultaneously its three main components (vehicle ownership, characteristics and use) and then deriving fuel demand from an identity such as expression (1).[7] The main advantage of this approach is to produce separate price and income effects for each fuel component while explicitly accounting for interactions between them. It is appropriate in the present framework since we are not only interested in producing predictions for the total amount of fuel consumed, but also for each of its components.

Disaggregated studies (e.g., see Mannering and Winston, 1985; Hensher and Milthorpe, 1987) suggest a strong relationship between ownership and use decisions implying that they should be modelled jointly. Wheaton (1982) and Hensher (1986) also acknowledge the importance of considering jointly fuel efficiency and distance travelled. Therefore, the three variables must be embedded in a simultaneous equations model which yields estimates of fuel consumption as a by-product. We now turn to examining how each component of the system may be modelled.

The first attempts to forecast the aggregate number of private-use vehicles consisted of extrapolations of historical trends based on some hypotheses regarding the evolution of the population. Later on, growth curve models (e.g. Whorf, 1973; Davis and Mogridge, 1976; and Tanner, 1978) aimed at predicting the vehicle ownership rate as a logistic function of time and a saturation level. Botton and Fowkes (1977) outline the drawbacks of this approach. Although Davis and Mogridge and Tanner use growth rate models, they even raise doubts about the existence of such saturation levels.

[7] Reza and Spiro (1979) undertook a slightly different approach by simultaneously modelling the demand for passenger cars, miles travelled, and car attributes, the latter being measured as the average weight of cars. The vehicle weight is often used as a proxy for its fuel efficiency, so the two approaches as very close to one another. Still, the fuel consumption is determined as the product of the number of vehicles, the distance they travel and their fuel consumption, the latter component being taken as given (exogenous).

Both Gallez (1994) and Jørgensen and Wentzel-Larsen (1990) underline the importance of including a lag structure in ownership models and the need to account for the saturation effect. Applying the stock-adjustment concept first introduced by Chow (1957) and Nerlove (1957) to the demand for durable goods such as private-use vehicles,[8] has provided an effective mechanism for determining the vehicle stock in a current period as a function of the sum of the previous stock, the incoming flow of new vehicle purchases, and the outgoing flow due to scrapage. Mellman (1975) and Ayres et al. (1976) provide surveys of these applications. Ben-Akiva, Manski and Sherman (1981) outline the main deficiencies of using the stock-adjustment approach for modelling vehicle ownership. Purvis (1994) also gives useful references to papers discussing aggregate automobile ownership modelling and saturation levels.

Train (1986), Mannering and Winston (1985), and Jørgensen and Wentzel-Larsen (1990) consider a discrete-continuous choice model system to estimate the joint ownership and use (distance travelled) decisions. The ownership equation is then used to calculate an inverse Mills ratio (Heckman, 1978) that is incorporated in the distance equation in order to correct for the selectivity bias due to limited availability of private-use vehicles. In these empirical studies, once other factors are accounted for, the selectivity term is insignificant at conventional levels. Hensher and Milthorpe (1987) further argue that if the interest is in predicting actual, as opposed to potential, distance travelled, Heckman's correction procedure is not necessary.[9]

[8] Among others, Jørgensen and Wentzel-Larsen (1990) acknowledge the importance of considering a vehicle as a durable good and its purchase as an investment.

[9] Based on Duan et al. (1984, p. 286), their argument is that when actual use is the main concern, the specification of the use equation can be limited to the randomly selected vehicle, but not on the use of other vehicles from the household's vehicle fleet, for instance.

Wheaton (1982) applies a similar method to aggregate cross-sectional data. He formulates a structural model in which the mileage equation is a function of the number of vehicles, their fuel efficiency, and some exogenous variables. Hypothesising that no contemporaneous correlation exists among the error terms of the three system equations, equation-by-equation OLS estimation may be undertaken. Wheaton tests this assumption by instrumenting out possibly endogenous variables and found no evidence of correlation based on such tests. While dealing with time series data, one must also account for the possibility of serial correlation (e.g., see Reza and Spiro, 1979).

In the context of aggregate simultaneous equations, each equation is usually specified in a linear or log-linear form. According to Wheaton (1982) and Dahl (1986), Box-Cox tests generally favour the log-linear specification for the stock and total distance equations, but the linear form for fuel consumption rates and average distance travelled. Each equation includes indicators of income and gasoline prices (Wheaton, 1982; Dahl, 1986; Dahl and Sterner, 1991). While prices of vehicles typically appear only in the stock equation, Wheaton also includes them in the fuel efficiency equation, in addition to an indicator for the level of urbanization. Wheaton and Hensher (1986) account for the potential impact of the vehicle fuel consumption rates on the distance they travel.

Both Gallez (1994) and Schimek (1996) find that income has a stronger effect on use than on ownership. De Jong (1990) distinguishes among fixed and variable costs associated with holding and using a private vehicle.[10] He finds that fixed costs are critical to the ownership decision, while variable costs mainly affect the intensity of use. Hensher and Milthorpe (1987) further differentiate fuel costs from other kilometre-dependent recurrent costs (discretionary maintenance and repair costs). These effects are not

[10] While this is a disaggregate cross-sectional analysis, fixed and variable costs are assumed to be the same for every household in order to produce homogenous aggregate costs. Furthermore, it is assumed that fixed costs only have an income effect.

separable, however, as they find that only recurrent costs are significant in the use equation, while fuel costs are relevant solely to the ownership decision.

Disaggregate studies of vehicle ownership (e.g. Ben-Akiva, Manski and Sherman, 1981; Mannering and Winston, 1985; De Jong, 1990; Prevedouros and Schoffer, 1992) also suggest that socio-economic and demographic factors such as the number of (members by) families or households, the population age and gender distribution, the unemployment rate, the number of driving licence owners and the population density in rural versus urban areas and land use might also have an effect on the private vehicle fleet size. Some of these analyses also highlight the importance of accounting for the fares and the quality of public transport services in this framework.

According to disaggregate studies such as those of Hensher (1986) and Hensher and Milthorpe (1987), the same kind of explanatory variables may have an impact on the distance travelled as well. Jørgensen and Wentzel-Larsen's paper (1990) further suggests that the proportion of business registrations might also matter via a substitution effect which operates between private- and commercial-use vehicles, the latest incurring no variable costs to the user.

More technical studies, such as the controlled experiment reported in Redsell, Lucas and Ashford (1993), reveal that the distance travelled, seasonal variations, temperature and other climatic conditions (see also the Organization for Economic Cooperation and Development, 1982, 1985; Chang et al., 1976; Eccleston and Hurn, 1979; Fwa and Ang, 1992) affect fuel consumption rates.

Dahl (1986) and Dahl and Sterner (1991) acknowledge the difficulty of obtaining accurate price and income elasticities with quarterly data. Their argument is that seasonal variations complicate matters and models based on such data can capture, at best, short-run elasticities, which are generally smaller than their long-run counterpart. However, the primary interest here is in obtaining good predictions for the three fuel consumption aggregates, not necessarily accurate individual estimates of the structural parameters.

Moreover, the iterative procedure, which we will make use of, produces the predicted values one period after another, so reliable short-run estimates are sufficient in the present context. While some studies dealing with quarterly data, such as that of Reza and Spiro (1979), apply seasonal adjustments to the data, our model must also account for quarterly variations in order to produce accurate predictions which vary from one quarter to the next. Hence, the model specification will account for seasonal variations by explicitly incorporating seasonal dummy variables and possibly other quarterly explanatory variables as well.

The simultaneous equations model proposed by Schimek (1996) seems the most appropriate to analyze jointly the three energy components. The following section proposes an econometric approach to estimate such a model and derive predictions of the dependent variables out of it.

2.5. Estimation and prediction techniques

A simultaneous equations model will thus be formulated to analyse the three energy components. The model specification will have to rely on external data sources, given the lack of survey data covering the intermediate period, between the two surveys. A dynamic structure will be embedded into the model in order to account for the trends and seasonal variations in the quarterly time series. The estimation technique will have to account for the relationships among the vehicle ownership, type and use decisions under the assumption of correlated errors. Furthermore, the estimation process must also account for the fact that the dependent variable series involve missing values for the intermediate period in order to produce predictions for the missing data.

2.5.1. The Bayesian approach

A natural way of dealing with the estimation problem is to use Bayesian sampling-based approaches. The main advantage of these approaches lies in their relative simplicity of application in the present context. In particular, they present an advantageous alternative to the classical econometric methods, such as standard maximum likelihood (ML) estimation techniques, which would have to be applied to a very complex recursive model in order to account for missing observations in the time series.

The Bayesian method includes, in addition to the empirical information comprised in the likelihood function, the investigator's experience and beliefs about the prior distribution of the parameters within the estimation process. Moreover, instead of restricting itself to some particular statistical characteristics of the parameters, which are of secondary importance, the Bayesian approach allows for the estimation of the parameter distribution (including the distribution of missing observations) in a completely natural and symmetric way.

Zellner (1971) summarises the earlier applications of the Bayesian method whose foundation lies on the Bayes formula which factorizes $p(\theta|Z)$, the posterior density of the parameters, θ, given the observations matrix, Z, as follows:

$$p(\theta \mid Z)=\frac{L(Z \mid \theta)p(\theta)}{p(Z)} \tag{2}$$

where $p(\theta)$ represents the prior density of θ, $L(Z|\theta)$ stands for likelihood function, and $p(Z)$ corresponds to a normalizing constant which insures that the posterior density integrates to unity.

Since the normalizing constant appearing in the denominator of formula (2) does not depend on the model parameters in any ways, it can be left out of the analysis so that the estimation process relies solely on the numerator's product:

$$p(\theta \mid Z) \propto L(Z \mid \theta)p(\theta) \tag{3}$$

The product of the likelihood times the prior is called the kernel of the posterior density and we say that the posterior density is proportional to this product.

Although the posterior density summarizes all the relevant information regarding the parameters and the data, it is sometimes required to obtain a single point estimate for each parameter of interest, on which one can condition on to take other decisions. This is accomplished by introducing a risk function that measures the cost of basing the decision on this point estimate rather than on the true value of the parameters. A Bayesian point estimate is then obtained by calculating the expected value of the risk function, given the posterior density of the parameters. For instance, under a quadratic risk function, a point estimate of a function, r(θ), of the model parameters is simply the expected value of this function over all admissible values of θ:

$$E[r(\theta) \mid Z] = \int r(\theta)p(\theta \mid Z)d\theta \tag{4}$$

2.5.2. Sampling-based techniques

In most applications, θ is a vector rather than a scalar and accordingly, the computation of the expectation (4) involves the evaluation of multidimensional integrals of complex expressions which do not have analytical solutions. Numerical integration (also known as Monte Carlo integration) may be applied in cases where the parameters have a proper posterior distribution function from which random variables can be easily sampled. Given $\{\theta^1,\ldots,\theta^n\}$, a random sample from the joint posterior density, the method produces an approximation of the desired expectation by replacing integrals by summations over the random draws:

$$E[r(\theta) | Z] \approx \frac{1}{n} \sum_{i=1}^{n} r(\theta^i) \qquad (5)$$

Whenever only the kernel of the parameter density function is known, however, more sophisticated techniques, such as importance-sampling (section 5.4 in Hammersley and Handscomb, 1964; Kloek and van Dijk, 1978; Rubin, 1987, 1988) or Monte Carlo Markov Chain (MCMC) approaches, are required. Sampling-based (or MCMC) techniques are iterative processes that allow one to generate random variates from the joint posterior distribution of the parameters by using simpler alternative distributions. Values so generated in each iteration are not independent, but rather form a Markov Chain.[11] The Bayesian estimate is then calculated as the means of the random variates generated in the last iterations. Usually, a certain number of iterations (called the burn-in period) are skipped in order to leave the algorithm some time to converge and hence reduce the influence of the choice of the starting values on the resulting estimates.

Under certain regularity conditions, the parameter estimator generated in this way is convergent. According to Stokey, Lucas and Prescott (1989), if the Markov chain is ergodic,[12] convergence occurs. Since the draws are not independent, an estimate of the variance of a parameter estimate cannot be calculated as the empirical variance of the values computed over the consecutive draws. Geweke (1992) provides methods to assess the accuracy of functions of the parameters based on spectral analysis. Standard

[11] The sequence $\{\theta^i\}$ of random variables form a Markov chain of order p if its conditional distribution can be written as follows: $p(\theta^i | \theta^{i-1},\ldots,\theta^0)=p(\theta^i | \theta^{i-1},\ldots,\theta^{i-p})$, for all $i>p$.

[12] A stationary sequence $\{\theta^i\}$ is ergodic if, for any two bounded mappings $Y:\Re^p \rightarrow \Re$ and $Z:\Re^q \rightarrow \Re$, $\lim_{n\rightarrow\infty} |E[Y(\theta^i,\ldots, \theta^{i+p})\, Z(\theta^{i+n},\ldots, \theta^{i+n+q})]| = |E[Y(\theta^i,\ldots, \theta^{i+p})]|\, |E[Z(\theta^i,\ldots, \theta^{i+q})]|$.

methods for estimating the variance of functions of the parameters under the posterior are proposed in Carlin and Thomas (2000, pp. 170-172).

Gordon and Bélanger (1996) and Gelfand and Smith (1990) summarize the main MCMC techniques among which the Gibbs sampler (Geman and Geman, 1984), constitute the simplest, and most widely used, representative. Starting from arbitrary values, $\theta=\theta^0$, of the parameters vector, $\theta=(\theta_1', \theta_2', \ldots, \theta_J')'$, the technique consists in generating, at the i^{th} iteration, a series of random variates, θ_j^i, from their conditional posterior distributions, given the most updated values of the other parameters:

$$\begin{array}{lll} \theta_1^i & \sim & p\left(\theta_1 \mid \theta_2^{i-1}, \theta_3^{i-1}, \theta_4^{i-1}, \ldots, \theta_J^{i-1}, Z\right) \\ \theta_2^i & \sim & p\left(\theta_2 \mid \theta_1^{i}, \theta_3^{i-1}, \theta_4^{i-1}, \ldots, \theta_J^{i-1}, Z\right) \\ \theta_3^i & \sim & p\left(\theta_3 \mid \theta_1^{i}, \theta_2^{i}, \theta_4^{i-1}, \ldots, \theta_J^{i-1}, Z\right) \\ \vdots & & \\ \theta_J^i & \sim & p\left(\theta_J \mid \theta_1^{i}, \theta_2^{i}, \theta_3^{i}, \ldots, \theta_{J-1}^{i}, Z\right) \end{array} \qquad (6)$$

The method is implementable and most efficient whenever the conditional posterior distributions have simple standard forms from which random variates can easily be drawn, while the joint posterior distribution is much more complicated.

Under mild regularity conditions, this algorithm yields values of the parameter vector such that if the series $\{\theta^i\}_{i=1}^{\infty} = \{(\theta_1^i{}', \theta_2^i{}', \ldots, \theta_J^i{}')\}_{i=1}^{\infty}$ converges after n iterations, then the statistics $(1/N)\sum_{i=n+1}^{n+N} r(\theta^i)$ will converge almost surely to $E[r(\theta)|Z]$, where $N>n$.[13] Roberts and Smith (1994) provide necessary regularity conditions for that type of MCMC technique to converge. According to Gordon and Bélanger, a sufficient (but not necessary) condition would be that the conditional posterior distributions, $p(\theta_j|\theta_{-j},Z)$ be positive for all values of θ_j, and for all $j=1,\ldots,J$.[14] In terms of prediction performance, the

[13] A sequence $\{\xi^i\}$ of vector-valued random variables ξ^i is said to converge almost surely to a limiting random variable ξ if $\Pr\{\lim_{n\to\infty}\xi^i=\xi\}=1$.

[14] The Gibbs sampling draws always form a Markov chain of first order. When conditional distributions are positive, the chain is also ergodic.

Gibbs sampler generally out performs the importance-sampling method and is less adversely affected by the model size.[15]

The Gibbs sampling method is particularly suitable in the present framework since it allows one to make use of the recursive structure of the model in the estimation process, rather than having to estimate all of the parameters simultaneously. This feature is especially desirable whenever it is combined with a data augmentation process, as discussed below. Although the Gibbs sampler is known to yield good results in practice (see, e.g., Gelfand and Smith, 1990; Kadiyala and Karlsson, 1997), it typically has a slow rate of convergence, especially whenever the draws are highly correlated across successive iterations. Geweke (1988) suggested a technique consisting in using antithetic draws that aims at improving its convergence rate. By drawing pairs of negatively correlated values of the parameters in each step, instead of randomly sampling from the conditional posterior distributions, the technique reduces the numerical variance of the estimates, hence accelerating convergence of the algorithm.

The Gibbs sampler may also be combined with other MCMC techniques such as the importance-sampling algorithm, the Metropolis-Hastings (M-H) algorithm (see Metropolis et al., 1953 and Hastings, 1970, for foundations but Chib and Greenberg, 1995, for an intuitive and understandable introductory treatment) or the data augmentation procedure (Tanner and Wong, 1987) if necessary. The intuition behind Tanner and Wong's data augmentation procedure consists in considering missing values of some random variable as additional parameters to be estimated. Together with the Gibbs sampling process, data augmentation enables, in each step of the iterative scheme, the computation of these additional values given the most updated values of the other parameters and the postulated formulation of the model. Once again, the recursive

[15] See Kadiyala and Karlsson (1997) for empirical comparisons.

structure of the procedure makes it relatively simple to implement. Just as for other parameters, estimated values for the missing observations can be derived by averaging the values computed in each steps, after having skipped a certain number of turns for the algorithm to yield convergent estimates. In our framework, predicted values of the dependent variables for the intermediary period can be derived in this way.

The last section describes in more details how these methods will be implemented in our empirical framework in order to produce the desired estimation and predictions. This will be done in Chapter 4. Chapter 5 will provide the empirical results.

2.5.3. Applications

With regard to prior distributions for the parameters, many choices are possible. An investigator may make the choice based on past experience and knowledge about the phenomenon under study, as well as on the ease of computations. Some priors are called uninformative because they impose the least restrictive structure on the parameter values. Uninformative priors have the main disadvantage of leading to improper posterior densities that do not integrate to unity, but rather diverge. Alternatively, informative priors may be specified in such a way that the posterior densities are well-defined. One special class of informative prior, the natural conjugate prior, is often used, in practice, in conjunction with MCMC techniques. This has the special feature of leading to conditional posterior distributions that have the same form as the postulated prior distributions. The application of the Gibbs sampling procedure, for instance, is thus simplified by postulating natural conjugate priors which have standard forms from which random variates can easily be drawn.

Along the lines of Schimek (1996), we postulate a simultaneous equations model for the three transport aggregates of interest: S, $\overline{D}$, and E. We assume a linear form for such a system, for reasons of convenience. Since our primary interest lies in obtaining predictions for the three dependent variables, we concentrate on the estimation of the unrestricted reduced form of the system, rather than on its structural form.

We have applied, in turn, the Gibbs sampling method with data augmentation to the following kinds of models in order to produce the predictions. First, assuming that the error terms of the three equations were independent from each other and from period to period, we considered a reduced form of the structural model. For each equation separately, a linear regression model, possibly involving lagged dependent variables to account for the serial correlation in the quarterly survey data, has been estimated. We hypothesized the Normal-Inverted-Gamma natural conjugate prior (Zellner, 1971) for the reduced-form model parameters and the parameter representing the variance of the error term, respectively.

The assumption of independence was then partly relaxed, allowing for an autocorrelation structure of order four (AR(4)) in the error terms of a given equation. Following Chib (1993) and Chib and Greenberg (1994), the additional autocorrelation parameters were assumed to follow a prior Normal distribution, truncated at the stationary region. Note that in order to include the initial observations in the estimation sample instead of conditioning on them, the autocorrelation parameters were computed via a M-H step.

Estimation of a simultaneous equations model has then been undertaken in order to allow for possible relationships among the equations. At first, a seemingly unrelated regression model (SUR) has been formulated (Zellner, 1971, Gordon and Bélanger, 1996). A natural conjugate prior for such a model involve a multivariate Normal distribution for the model parameters, and an Inverted-Wishart distribution for the variance-covariance matrix of the error terms. This combination of priors is simply an extension of the previous priors to the multivariate case.

We have also estimated a pure vector autoregression (VAR) model involving only, in the specification of each equation, lags of the three dependent variables and deterministic explanatory variables (namely, a constant and seasonal dummy variables). Litterman (1986) favours Bayesian estimation of VAR models over classical estimation of structural econometric models for forecasting purposes. Kadiyala and Karlsson (1997) contrast prediction performances reached with different MCMC methods and prior choices in this framework. Finally, we have generalized Kadiyala and Karlsson's method to handle stochastic explanatory variables in addition to deterministic variables.

The comparative advantages of the VAR model setting are many. In both classical and Bayesian analysis, it is easier to estimate than the competing simultaneous equations or the autoregressive moving average (ARMA) models (Kadiyala and Karlsson, 1993). In particular, it avoids the identification problems which may arise due to parameter equality restrictions, on the one hand, and the necessity of choosing the exact number of autoregressive and moving average components to include into the model, on the other hand. Furthermore, a Bayesian treatment of the VAR model circumvents the degrees-of-freedom difficulty frequently faced by the classical approach.[16]

In all cases, the methods described in the literature had to be generalized to include a data augmentation stage aimed at computing the predicted values of the dependent variables for the intervening period. Whenever the econometric model involves lagged dependent variables as regressors, this complicates matters since the

[16] Litterman (1986) argues that in the Bayesian specification framework, there is no trade-off between decreasing bias (by adding more explanatory variables) and increasing variance (by narrowing the degrees of freedom margin, given the limited number of observations) because the loss function is minimized by including all relevant variables along with their priors. Although he acknowledges there is a practical limit to the number of such variables, he stresses that this limit is dictated by computational feasibility, but not by the lack of degrees of freedom. Kadiyala and Karlsson (1993) go along with Litterman in saying that the Bayesian approach handles large VAR models more easily than classical econometric methods.

predictions obtained for a given period will enter as predictors. This is the chain rule of forecasting described by Litterman (1986). It is thus to be expected that predictions at the beginning of the prediction period will be more precise and accurate than those obtained at the end of it because the former ones will rely on observed lagged values of the dependent variables, while the latter ones will be constructed from predicted values of the dependent variables. The very last predictions will thus involve a building-up of the prediction errors, summing those of past predictions in addition to the current one.

For a stationary ergodic process, however, the effect of past values on future ones decreases with the time interval between them. Therefore, recent prediction errors should have a larger impact on current predictions than more dispersed ones, and so their sum should be bounded above. We thus expect predictions at both ends of the prediction spell to be more accurate than those in the middle of the interval between the two surveys.

Chapter 6 will compare the results obtained from the estimation of alternative model formulations in order to determine the one yielding the most reliable and accurate predictions. It will also question the underlying assumptions.

Chapter 3

Data Processing

This chapter describes the preliminary stage of data processing performed on survey-based data. Recall that we are interested in deriving compatible estimates for the number of private-use vehicles, the average distance they travel, and their weighted fuel consumption rates from both surveys. Separate estimates are produced for cars and for light trucks and vans. These incomplete time series will then be used, in subsequent chapters, as dependent variables for an energy demand simultaneous equations model in order to predict values for the intermediary period between the two surveys.

The next section describes the surveys in detail. Section 3.2 explains how the variables of interest are calculated for this study. The preliminary stage to data processing consisting of imputing a fuel consumption to some observations for which it cannot be calculated directly from the answers to the survey questionnaires is described in Section 3.3. Section 3.4 identifies the main differences between the two surveys. The two following sections propose a series of adjustments that were applied to the raw estimates from each survey in order to account for those differences and produce compatible estimates.

3.1. Survey data

Following the Second Oil Crisis, Statistics Canada (StatCan) undertook the FCS for the account of Transport Canada. To perform this survey, which covers a period of nearly ten years, StatCan randomly drew samples from the provincial registration files[1] that already provide detailed information on the features of selected vehicles. A diary was then sent by mail to the owner of the sampled vehicle in which he was responsible for reporting all fuel purchases performed for this vehicle during a full month, flags indicating if these sufficed to fill in the gas tank, the purchasing dates, the unit fuel prices,

[1] A more global study of the Canadian private transport sector would have to include both Norther Territories and incoming traffic and fuel purchases from across borders (in particular, from the U.S.).

the type of fuel purchased, and the number of kilometres written on the odometer, at the same time.

The FCS covers the period from the fourth quarter of 1979 to the last quarter of 1988 for cars, while data on light trucks and vans were only collected from the fourth quarter of 1981 to the last quarter of 1987. The survey was then interrupted as, as reported StatCan in its first report on the NaPVUS, "the oil crisis had passed and concerns about petroleum reserves had been further mitigated by new oil discoveries, new oil extraction technologies and more fuel efficient vehicles". Preoccupied by the new environmental issues of fuel pollutant emissions, the Canadian Minister of Natural Resources (NRCan) requested that StatCan undertake the NaPVUS. This survey was designed to be an improved version of the FCS, and to be undertaken under sequential episodes, rather than on a regular basis. It took place from the last quarter of 1994 to the third quarter of 1996, and is expected to be repeated soon.

The NaPVUS sampling unit is the household, randomly drawn the *Labour Force Survey* (LFS) sample. The characteristics of the household, its vehicle fleet, and the general use made out of it were first collected via a *Computer Assisted Telephone Interview* (CATI). A vehicle was randomly selected among the household's vehicle fleet for completing the second stage of the survey. Respondents were then sent by mail a fuel purchase diary similar to the FCS one. All users of the selected vehicle were requested to report each fuel purchase dedicated to this vehicle during a whole month, as well as other relevant information at the moment of the purchase. Consequently, the NaPVUS does not only provide information on the selected vehicle, as does the FCS, but also on its drivers' profile.

3.2. Estimation of survey-based variables

We distinguish between a fuel purchase, that does not fill in the gas tank, and a fill-up, which does. It takes at least two fuel fill-ups reported in the diary during the survey month to obtain an estimate of the selected vehicle fuel consumption. From the date of the first fill-up to that of the second one, the fuel consumption can be calculated: it is equivalent to the amount of fuel purchased the second time.[2] Of course, if more than two fill-ups occurred during the survey month, the amounts of fuel purchased each time since the first fill-up are cumulated. The fuel consumption for the period between the first fill-up to the last one is then extended to the whole month by assuming that it is representative of the monthly fuel consumption.

Similarly, two fuel purchases realized over the survey month are sufficient to infer the distance travelled by the selected vehicle from the corresponding odometer readings. The difference between the last odometer reading and the first one yields the exact distance travelled by the sampled vehicle during the period.[3] The distance travelled during the survey month is then estimated on the basis of the assumption that the average distance travelled between the first and last odometer readings is representative of the average distance travelled during the month. The daily average distance travelled is thus multiplied by the total number of days comprised in the survey month.

In order to compute the fuel consumption rate of a selected vehicle involving at least two fill-ups during the survey month, we simply divide the estimate of the monthly

[2] There is no way to figure out the exact fuel consumption from less than two fill-ups. If the tank was not full at first, the quantity of fuel necessary to fill it in exceeds fuel consumption. On the contrary, even if the tank was full initially, the amount of fuel purchased will be inferior to the quantity consumed.

[3] For the FCS, additional odometer readings were requested at the exact moment the diary was received and returned. Therefore, distance estimates should be generally more precise in the FCS than in the NaPVUS because they are based on longer observation periods. In terms of adjusting the NaPVUS distance estimates, however, not much can be done about this.

fuel consumption (expressed in litres) by the estimated total distance travelled during the month (expressed in hundreds of kilometres) as derived above. Note that even if only two fuel purchases are sufficient to estimate the distance travelled by the sampled vehicle during the survey month,[4] two fuel fill-ups are necessary to estimate its fuel consumption and fuel efficiency.

3.3. Imputation of incomplete diaries

For the NaPVUS, which involves relatively small sample sizes, it became important to retrieve as many fuel purchase diaries as possible for estimation purposes. In order to do so, a fuel consumption has been imputed to sampled vehicles for which less than two fuel fill-ups, but at least two fuel purchases, were registered during the survey month. This section describes this preliminary stage of data processing which has been performed by StatCan, but to which the Automobile Mobility Data Compendium (AMDC) actively took part as an advisor for the NaPVUS.

Note first that when the news release described in Section 2.2.1 of Chapter 2 was issued, imputation for the first-quarter NaPVUS data had not been undertaken yet. Ignoring NaPVUS vehicles with less than two fill-ups but more than one fuel purchase was partly responsible for the huge discrepancies that were initially observed between the FCS and the NaPVUS estimates. As those vehicles, used relatively less intensively, were excluded from the NaPVUS estimates but not from the FCS ones, it biased the NaPVUS estimates of the average fuel consumption per vehicle upwards. The AMDC highlighted this fact and recommended the adoption of the imputation procedure for the NaPVUS.

[4] Two fuel purchases are necessary to derive NaPVUS distance estimates. For the FCS, the distance can always be calculated for reasons outlined in the preceding footnote.

A prediction model for fuel consumption (in logarithmic form) has been formulated. This has been estimated on the basis of observations for which fuel consumption could be estimated using the method described in the preceding section. The estimates were then used to predict, or impute, a fuel consumption to vehicles with less than two fill-ups, but more than one fuel purchase. The fuel consumption rates of such vehicles have then been determined based on the predicted fuel consumption and the distance estimate.

The percentage of incomplete diaries varies, for the FCS, from one quarter to the next. It reaches up to 40 percentage points in the second quarter of 1985. For the FCS, the prediction model for the imputation stage is a log-linear model involving the natural logarithms of the distance travelled and the vehicle weight as explanatory variables.[5]

The highest percentage of imputed vehicles for the NaPVUS was 18.6 and it occurred in the first quarter the survey was undertaken. Based on experiences performed on the FCS data, the NaPVUS model retained for imputation purposes also has the log-linear form, although it involves a larger range of explanatory variables. Dummy variables for the number of cylinders of the selected vehicle enters the model as a proxy for the vehicle weight which is not reported in the NaPVUS, while it was given in the registration files from which the FCS samples were drawn. The prediction model also involves the natural logarithm of distance travelled, the number of fuel purchases done during the survey month, the household size, and dummy variables for the vehicle type

[5] Quarterly reports on the FCS (1979-1989) do not explicitly describe the imputation model, although they note that the fuel consumption was estimated on the basis of the distance and the vehicle weight whenever it could not be determined directly from at least two fill-ups. The annual report for October 1979 to 1980, however, presents some exploratory regression analyses that yielded to the imputation model. Since this model corresponds to the one suggested in the quarterly reports, we assume it is the one considered throughout the survey period.

(car versus light truck/van), the transmission type, and the household region of residence (rural versus urban).

The AMDC's task consisted in helping StatCan develop a prediction model. Throughout the imputation stage, the AMDC closely followed StatCan's work, issuing, whenever necessary, a series of comments and suggestions regarding special data features, relevant explanatory variables to be tested in the prediction model, alternative functional forms to model fuel consumption, and diagnostic tests for the model specification. Bonin and Boucher's report (1997) summarises the AMDC recommendations regarding the NaPVUS imputation.

The imputation procedure for the NaPVUS is currently being revisited by Sylvie Bonin, chief analyst at the AMDC (see Bonin, 1999). The consequences of having ignored some of the AMDC's recommendations for respecting the time schedule and monetary constraints will be assessed. The merger of the NaPVUS to the *Vehicle Fuel Economy and Emissions System* (VFEES) and the *Vehicle Fuel Economy Information System* (VFEIS) microdata banks (see Boucher, 1999) will also allow us to consider a wider range of explanatory variables. In particular, the explanatory power of a categorical variable for the vehicle weight and its average lab-tested fuel efficiency will be tested. Furthermore, the consequences for compatibility of using different imputation specifications for successive stages of the NaPVUS will be analysed.

3.4. Major differences between the two surveys

Because of the specifics, to be outlined below, of certain questionnaire items and survey fields, it is risky to make a direct comparison between estimates derived on the basis of the data from the two surveys. Not having taken them into account in comparisons produced in NRCan's news release was also responsible for the large discrepancies between the estimates produced from both survey data. In this section we will summarise the five major differences between the surveys. Differences that come into play in our discussions in the next two sections.

The main difference lies in the sampling unit itself. The base sampling unit for the FCS is the vehicle, drawn from provincial vehicle registration files, whereas the NaPVUS sampling unit is the household, selected from the LFS samples. Consequently, the NaPVUS questionnaire is much more detailed and collects information on household characteristics as well as data on vehicle use. Furthermore, the difference in sampling units leads to several discrepancies in survey fields that impact significantly on comparisons. These differences will be analyzed in detail in the next two subsections. For now, it is sufficient to say that administrative delays in updating vehicle registration files resulted in the exclusion of certain vehicle categories from the FCS sampling field.

A second distinction is the categorization of vehicle types. The FCS only distinguishes cars from light trucks and vans. In particular, it does not allow one to differentiate minivans from light trucks. On the contrary, the NaPVUS vehicle categories are much more detailed and distinguish among cars, minivans, light trucks, and vans.

A third difference is the way unused vehicles during the survey month were processed. They were saved in the NaPVUS microdata bank but not in that of the FCS. This introduces a selection bias into the FCS estimates. As a result, the number of private-use vehicles in Canada was underestimated for the FCS sampling period. And, consequently, the average per vehicle distance and fuel consumption estimates for the FCS were overestimated.

A fourth difference is that vehicles registered as commercial vehicles are not part of the FCS sampling field but are included in the NaPVUS as long as they were also used for private purposes.

Finally, another difference lies in that vehicles on a long-term lease and vehicles provided by the employer of one of the household members were considered in the NaPVUS but not in the FCS, which only surveyed privately owned vehicles.

In the next two sections, we propose a way of accounting for these differences in order to generate compatible time series on the basis of data from the two surveys. Estimates for the variables of interest are derived from the microdata banks of the two surveys. The degree of aggregation of the survey data is set at the quarterly level for a reason that will be explained in detail below. In particular, quarterly estimates of the average number of vehicles, total distance travelled, and fuel consumption are required for the purposes of our study. The estimates will serve to generate complete time series for the number of vehicles, average distance, and the weighted fuel consumption rate (in litres per 100 kilometres) for each type of private-use vehicles — cars and light trucks/vans. All adjustments to survey-based estimates are described in more detail in Appendix A.

3.5. NaPVUS estimates

This section summarises the data processing leading to the NaPVUS estimates used in this study.

3.5.1. Raw NaPVUS estimates

There are two types of statistical weights in the NaPVUS microdata bank. This is a direct consequence of the underlying survey method. First, the households, which were randomly selected from the LFS samples, were contacted in a telephone interview. Those who agreed to participate in the second stage of the survey received a diary by mail for reporting all fuel purchases during the survey month for one of their vehicles, which was randomly selected from the household vehicle fleet established at the time of the telephone interview. The first statistical weight, to which we will refer as the background questionnaire weight (BQW), is determined on the basis of the telephone interview and is assigned to the household to reflect the degree to which it is representative of the Canadian population. The second statistical weight, related to the fuel purchase diary, or diary weight (DW), reflects the degree to which the selected vehicle is representative of the whole fleet of private-use vehicles in the Canadian provinces.

Sample sizes for both surveys are reported in Tables A.1 and A.2 of Appendix A. Note that the number of vehicles selected to fill out the FCS fuel purchase diary has been considerably reduced starting from 1986. Sample sizes have been cut to nearly one third of their previous levels. Based on sample sizes, estimates are thus likely to be less precise at the end of the survey period. Furthermore, as Section 3.6.3 will show, the fact that a narrower variety of private-use vehicles are represented in the sampling field for that period also contributes to a lack of precision.

In terms of selected vehicles, the average sample size for the FCS is more than twice as large as that for the NaPVUS. Despite this fact, the NaPVUS estimates may still be better than the FCS ones because they are based on a more extended and complete sampling field. Indeed, Section 3.6 will provide a list of vehicle categories that were excluded from the FCS samples, due to restrictions imposed by the sampling basis (registration files). Compared to the NaPVUS, the FCS over-sampled light trucks and vans (on average, around 40 percent of the NaPVUS vehicle sample is composed of light trucks and vans, while more than 65 percent of the FCS vehicles belong to the same vehicle class). This compensates for the fact that they represented a smaller proportion of the light vehicle fleet during the first survey period.

Since the NaPVUS respondents who filled in the fuel purchase diary is a sub-sample of the households who participated to the first-stage telephone interview, the BQWs provide better estimates than the DWs. StatCan, which conducted the survey, has generated its estimates for the main variables of interest relating to private-use vehicles by applying the DWs. It was justified because most of these variables were derived from the diary. However, in order to depict a more accurate picture of the private-use vehicle fleet in Canadian provinces, we consider a new approach to estimation based on the BQWs.

Specifically, the private vehicle fleet size can be calculated in two ways: either by using information on all vehicles provided by the household at the time of the interview, or by using information from the fuel purchase diary for the vehicle selected to be part of the second stage of the survey. We adopted the first method. Estimates of average distance and average fuel consumption are based on diary information, because the telephone interview did not provide accurate data for this purpose. Assuming that the average distance and fuel consumption performed by households agreeing to participate in the second stage of the survey do not differ substantially from those of households who answered the telephone interview, total estimates for these two variables are obtained by multiplying average values by the number of vehicles determined by means of the interview.[6]

3.5.2. Adjustment for the non-response to certain questions

Because the telephone interview respondents in the first stage of the NaPVUS survey were required to provide information on all vehicles used by the household, some data may be missing. For instance, the vehicle type was not identified in the case of some vehicles selected for the purpose of filling out the fuel purchase diary. As a result, there is a problem when disaggregation by vehicle type is required. By way of illustration,

[6] To confirm this hypothesis, an analysis of the NaPVUS non-responses and partial non-responses is needed. Given the lack of detailed information on the characteristics of respondents and non-respondents to the first or second stages of the survey, this approach seems reasonable.

we obtain a higher number if we estimate the number of vehicles on the basis of the NaPVUS sample than if we calculate the sum of estimates by vehicle type (cars and light trucks/vans treated separately). The difference between the two total estimates may be attributed to vehicles of unidentified type.

Since cars and light trucks and vans have very different characteristics, in terms both of use and of the degree to which they are representative of all vehicles, they are seldom grouped together in empirical studies. A comparative study by vehicle type based on the NaPVUS-derived estimates would underestimate not only the total number of vehicles but also all other overall quantities, such as total distance travelled and fuel consumed by all vehicles. Therefore, an adjustment is required for estimates by vehicle type so that they better reflect overall use of private vehicles in Canada.

Moreover, for time series modelling such as this, an adjustment of this kind will allow one to reduce seasonal variations due to non-response. Otherwise, if there were more missing values for one of the vehicle types during a given quarter, the size of the vehicle fleet, the total distance travelled and the total fuel consumption would be significantly underestimated for that quarter. The lower figures would be due not to seasonal factors but to the greater number of missing values for this quarter. The proposed adjustment, which happens on a seasonal frequency, will permit us to limit additional biases in the time series of interest, and the econometric model will thus be able to reproduce more accurately actual seasonal variations during the prediction period.

We therefore propose a distribution of vehicles of unidentified type as well as use-related variables across the raw estimates by vehicle type for the NaPVUS identified-type vehicles, based on the latter initial repartition. In order to remain consistent with the initial distribution of the identified-type vehicles, the difference between overall estimates for all vehicles and the sum of estimates by vehicle type is distributed between the two types proportionally to this distribution. Appendix A contains a detailed description of the adjustment method. The estimates required for calculating the proportions referred to in the next section, which deals with adjustments to the FCS estimates; will have gone through a similar adjustment.

3.5.3. Categorization of minivans

We should also draw attention to another feature of the surveys that must be factored into any comparison: the processing of minivan data. Minivans are closer to light trucks in shape but closer to cars in terms of use. Their categorisation is therefore a difficult issue.

Minivans have been assigned to the light truck class for the purposes of this study so that both vehicle types can be more easily compared with those in the FCS, which groups together minivans and light trucks. It should be noted however that this will introduce a bias into the NaPVUS estimates since the minivan class was associated with cars in the specification for the imputation model discussed in Section 3.3. The revision of the NaPVUS imputation stage should allow us to avoid such a problem since it will propose a prediction model in which the categorical variables for the vehicle type will differentiate minivans from car and light truck/van classes.

3.5.4. Vehicles used primarily for commercial purposes

As mentioned above, the NaPVUS data incorporate a large proportion of commercial-use vehicles because any commercial-use vehicle which is also used for private purposes is part of the survey. Note that there is no way of identifying vehicles with commercial registrations in the databank. However, it is possible to determine whether a selected vehicle was solely private-used by appealing to a variable derived from the survey background questionnaire. The percentage of time a vehicle was exploited for commercial purpose, if any, during the month preceding the telephone interview is also part of the background questionnaire. Furthermore, answers to the first stage interview allow us to identify the owner of the selected vehicle: whether the vehicle was rented on a long term lease, or owned by the household itself, or by an employer of a household member.

There is no clear-cut way to distinguish between private and commercial vehicles unless these vehicles are solely dedicated to one use or another. Among vehicles used for both motives, the distinction is somewhat arbitrary. We will divide vehicles according to their primary use. It is natural to assume that vehicles registered as commercial vehicles are mainly used for commercial purposes. Since the FCS sampling field excludes such vehicles, primarily commercial-use vehicles should be more numerous in the NaPVUS samples than in the FCS ones.

Assuming that the percentage of time the selected vehicle was used for commercial purposes during the month preceding the NaPVUS first stage interview is representative of its general use, one can base the vehicles categorization according to their primary use on this variable. Figure A.1 of Appendix A illustrates a typical empirical distribution of this variable, based on the NaPVUS data for the fourth quarter of 1994, for the whole vehicles partly used for commercial purposes, as well as for cars and light trucks and vans separately. Although based on a relatively small number of observations (246 vehicles were partly used for commercial purposes during this quarter), the typical empirical distribution shows three distinct modes around 10, 50, and between 80 and 90 percent of commercial use, respectively.

Given that the NaPVUS sampling field may involve a share of vehicles registered as commercial ones, a conservative criterion has been retained to differentiate vehicles according to their primary use. Specifically, it was agreed that a minimum of 75 percent of commercial use was a reasonable threshold for classification in the primarily commercial-use vehicle group. This criterion was initially established on the basis of estimates derived from the non-imputed NaPVUS data for the fourth quarter of 1994 but proved to be appropriate as well for subsequent quarters and after completion of the imputation stage.[7]

3.5.5. Not privately-owned vehicles

Initially, an adjustment was made to the FCS estimates for employer-owned vehicles, but they are now considered primarily commercial-use vehicles for obvious reasons. It is highly probable that they are registered as commercial vehicles, and they will therefore be processed in the same way as the NaPVUS vehicles with a commercial use rate of at least 75 percentage points. In other word, they are excluded from the empirical analysis.

Moreover, vehicles on a long-term lease are now considered private vehicles. One reason is that the long-term leasing is similar to the purchase of a new vehicle for private purposes: the use of a vehicle on a long-term lease resembles that of a privately owned vehicle.[8] Furthermore, an adjustment to the FCS estimates on the basis of the proportion of leased vehicles in the NaPVUS, such as the one that was initially considered, overestimated the representation of leased vehicles in the first survey, because the popularity of long-term leasing has grown significantly in recent years. The number of leased vehicles as a percentage of the total private vehicle fleet should

[7] See Appendix A and Boucher (1998a) for a justification of the choice of the partitioning criterion. Complementary analyses (see Boucher, 2000, and Boucher and Bonin, 2000) have shown that predictions do not change that much when primarily commercial-use vehicles are kept in the NaPVUS data sets. They follow very similar trends and seasonal variations, but their precision is adversely affected by the difficulty to identify relevant explanatory variables to capture the commercial share in the data.

[8] In fact, since vehicles on a long-term lease are generally returned to the dealer after a couple of years, they represent a relatively younger vintage, and accordingly, might be used relatively more intensively than the overall privately-owned vehicle fleet (unless there is a kilometres limit in one of the contract clauses). However, the long-term leasing market, just as the market for new vehicles, attracts consumers who like to have their vehicle under warranty and use it relatively intensively. This kind of consumers existed in the 1980's, as in the 1990's. Therefore, we feel that excluding vehicles on a long-term lease from the NaPVUS data would further biase our results.

therefore have been lower in the eighties than in the nineties. So by including vehicles on a long-term lease in the private vehicle class, we avoid having to make an adjustment that would inevitably build an upward bias into the FCS estimates.

3.6. FCS estimates

A number of adjustments are required to ensure that the FCS and NaPVUS estimates are comparable. Most of the adjustments are due to the FCS sampling unit, while some are related to the fact that the NaPVUS sampling field is more extended.

3.6.1. Adjustment for vehicles excluded from the sampling field

As mentioned above, the FCS sampling unit is the vehicle, drawn from provincial vehicle registration files. Since the files had not necessarily been updated at the time of the sampling for the survey, which was repeated on a monthly basis, it was impossible to include some vehicle categories for certain provinces in the sampling during given quarters. Table A.3 of Appendix A summarises the categories of vehicles that were excluded for each FCS quarter. Since these are specific vehicle categories, their exclusion would result in biased estimates derived from the survey data.

The first FCS adjustment is based directly on StatCan estimates, as reported in its quarterly FCS reports (1977-1988). The reports contain revised figures that cannot be reproduced from the microdata bank. We must therefore conclude that StatCan made adjustments for vehicle categories not included in the registration files of some provinces at the time of sampling. Comparative analysis of the StatCan quarterly results and estimates derived from the microdata bank enabled us to deduce the correction factors used by the agency to compensate for the incomplete information for each variable of interest. Specifically, correction factors, applied to the number of vehicles, the total distance and the total fuel consumption, were obtained separately for cars and for light trucks and vans. Average distance and weighted fuel consumption rates can be derived from these variables by calculating the ratio of total distance to number of vehicles, and the ratio of total fuel consumption to total distance, respectively.

The need for such an adjustment determined the level of data aggregation. Since the StatCan estimates are published only every quarter, adjusting to a more disaggregated level would have been complicated. The databank provides monthly estimates, but no specific rule is available for disaggregating the correction factor for missing vehicles at this level. Furthermore, StatCan's method of adjustment was not explained in its reports, so any attempt to come up with a new method would inevitably have yielded different results from those of the agency.

Note that an adjustment to bring the sum of estimates by vehicle type in line with the overall estimates for all vehicles, as was done for the NaPVUS estimates, would be unnecessary in the case of the FCS. The FCS vehicles are drawn from registration files, so this type of information is available. The difference between the overall estimate and the sum of the estimates by category is always zero in the FCS throughout the sampling period.

The next three adjustments are based on the distribution of the NaPVUS data for each quarter. Assuming that this distribution also applies to the years in which the FCS was conducted, it is possible to make at least a partial correction to the FCS-derived estimates to account for the exclusion of certain vehicle categories from the sampling field. However, the corrections are approximated because the distribution of all vehicles and the use of private vehicles have changed over the last twenty years. As argued in Appendix A, such adjustments, although imperfect, are nevertheless preferred to no adjustment at all.

3.6.2. Adjustment for unused vehicles

Vehicles that were not driven during the survey month were sampled in the FCS, but the information was not saved in the microdata bank. As a result, the total number of vehicles for each survey quarter is underestimated. Similarly, while estimates of distance and fuel consumption attributable to all vehicles are unaffected by the exclusion of unused vehicles, which yield nil values for the two variables, the average values of the variables are overestimated because they are based on a restricted number of vehicles.

The bias caused by the exclusion of vehicles that were not used during the survey month can be corrected only partially through a second adjustment. Based on the assumption that the percentage of non-driven vehicles during the FCS and the NaPVUS sampling periods was the same, estimates of the vehicle fleet generated from the FCS data can be corrected using the NaPVUS information. The percentage of leased or privately owned vehicles not driven during the NaPVUS survey can be extrapolated to the FCS data.

Thus, the FCS estimates can be increased by the above percentage of the total vehicle stock. However, since private vehicle use has steadily grown over the last twenty years,[9] the number of vehicles not driven for a whole month is likely to be higher for the FCS than for the NaPVUS. Accordingly, the proposed adjustment is a minimal correction of the bias caused by the destruction of the FCS observations on unused vehicles. At the same time, estimates of total distance and fuel consumption do not require such an adjustment because they are in no way affected by the exclusion of unused vehicles, which register nil values for these variables.

[9] A related AMDC report (see Boucher, 1998b) has shown that while the number of vehicles has grown over the last twenty years, private vehicles have also been used more intensively since the early nineties.

In the past, unused NaPVUS vehicles were identified as vehicles showing zero distance travelled during the survey month. However, it appears that vehicles to which no fuel consumption could be attributed because they displayed less than two fuel purchases and two fill-ups were assigned zero distance too in the NaPVUS data bank.[10] Thus, using the zero distance criterion to estimate the number of non-driven vehicles in the NaPVUS would produce an overestimate.

In addition, since the distance travelled by the selected vehicle is a diary-based variable, it encompasses many missing values owing to the partial non-response bias. Specifically, the variable is missing for all households that participated in the telephone interview but refused or failed to fill out the diary. Thus the new adjustment, designed to distribute the missing values among nil values and values above zero, would inflate the unused vehicle estimates even more.

For both these reasons, we established a new criterion for calculating the number of non-driven vehicles during a given month. It relies on a telephone interview-based variable --- the dummy variable indicating whether the selected vehicle was used during the thirty days preceding the interview --- and involves fewer missing values. The distribution of missing values between vehicles driven and vehicles not driven during a month will thus generate a more reliable and lower estimate of the proportion of non-driven vehicles.

Furthermore, this criterion is less vulnerable to non-response bias problems. For instance, a respondent could state that he did not use the selected vehicle during the survey month so as to avoid having to fill out the fuel purchase diary. On the other hand,

[10] This also implies that average distance and average fuel consumption are overestimated, because these vehicles registering lower-than-average values for the two variables are not represented. Nonetheless, this does not cause further disparity between estimates for the two surveys, because the FCS vehicles presenting similar characteristics were processed in the same way.

a respondent does not have a strong motive for giving a negative answer to the question concerning the use of the selected vehicle during the month preceding the telephone interview. At most, his incentive would be to avoid having to answer a few additional questions. The adjustment for exclusion of unused FCS vehicles is therefore based on this new criterion.

Applying the criterion, estimates of the percentage of leased or privately owned vehicles not driven during the month prior to the NaPVUS telephone interview have been derived for each quarter. All correction factors based on quarterly proportions derived from the NaPVUS data were calculated in the same way. The NaPVUS estimates for the same quarter were matched in pairs to generate quarterly correction factors based on the largest possible number of observations and thereby increase their level of significance. Given that the NaPVUS sampling period as a whole covers two full years, each adjustment factor is based on data from the two corresponding quarters.[11]

3.6.3. Adjustment for new vehicles

A close examination of the FCS reports revealed that certain classes of new vehicles were not included in any provincial registration files for some quarters. By new vehicles, we mean those manufactured during the survey year and those whose model year was the following year but which were put on the market during the current year. All these vehicles were excluded from the FCS sampling field during the 1984-89 period. For 1979-83, on the other hand, vehicles whose model year was the current year were sampled starting in the fourth quarter of that year.

[11] For example, the adjustment factor applicable to the first quarter data is calculated on the basis of the NaPVUS observations in the first quarter of 1995 and the first quarter of 1996, and the adjustment factor applicable to the fourth quarter data is calculated on the basis of observations in the last quarter of 1994 and the last quarter of 1995. The adjustment factors proposed below to compensate for the exclusion of other vehicle categories from the FCS sampling field are calculated in a similar fashion.

Thus vehicles whose model year was the following year but which were put on the market toward the end of the current year were always systematically excluded from provincial registration files. However, assuming that these vehicles came on the market only in the fourth quarter of any given year,[12] only an adjustment to the FCS fourth-quarter estimates is required. The adjustment is combined with a correction to compensate for a lack of current-year vehicles if they too were excluded from the registration files. In short, three kinds of adjustments are required to compensate for the exclusion of new vehicles from the FCS:

a) For the first three quarters of each year, the correction is based on the proportion of vehicles of the current model year in primarily private-use NaPVUS vehicles;

b) For the fourth quarters during the 1979-83 period, the adjustment is based solely on the proportion of vehicles of the following model year put on the market during the fourth quarters of the NaPVUS. Note that this percentage is very small and cannot be considered significant because its numerator is based on a very small number of observations;[13]

c) For the fourth quarters of the1984-89 period, the adjustment is based on the cumulative proportion of current- and next-model year vehicles in NaPVUS estimates.

[12] One can assume that this applies to most vehicles of the following model year. Indeed, in all NaPVUS samples, no vehicles of the following model year appear until the fourth quarter of any given year. In addition, since new vehicles are coming out earlier and earlier each year, they are even less likely to come out before the fourth quarter of an FCS survey year than during the NaPVUS sampling period.

[13] Fewer than thirty observations for the two fourth quarters of NaPVUS as a whole. Yet StatCan recommends that estimates based on such a small number of observations not be published. That being said, NRCan wanted this adjustment, however small the proportion was. Our view is that application of this adjustment would in any case have only a marginal impact on the final estimates.

3.6.4. Attempts to retrieve primarily commercial-use vehicles

Since vehicles with a commercial registration are excluded from the FCS, the survey involves a smaller share of primarily commercial-use vehicles than the NaPVUS. Nonetheless, several attempts to isolate such vehicles from the FCS databank have been undertaken. Since the FCS only provides information on the vehicle characteristics, but not on its user, there is no way to infer which use the selected vehicle was primarily dedicated to. As for the two preceding adjustment procedures, the natural way to proceed is to rely on relevant information from the NaPVUS. But since the NaPVUS does not provide information about the selected vehicle registration class, the adjustment cannot be based on the proportion of primarily commercial-use vehicles apart from those with commercial registrations.

The general idea underlying all of the partitioning trials was to extrapolate the NaPVUS percentage of commercial-use vehicles to the FCS period. Different formulations of discrete choice models have been explored to model either the incidence of using a vehicle primarily for commercial purposes, or the proportion of commercial use itself.[14] The models were estimated using the NaPVUS sample, and then used to predict the dependent variable based on the FCS information. Given the limits of the information provided by the FCS, the set of admissible explanatory variables for such models had to be restricted to the vehicle characteristics. Some household characteristics proved to have an important explanatory power in the prediction model, and so the different trials were meant to fail.

Therefore, we were unable to retrieve the remaining primarily commercial share of vehicles from the FCS data. Since it already excludes vehicles with commercial

[14] See Boucher (1998c) for more details on the different inference trials.

registrations and involves only privately owned vehicles, the FCS sampling field however excludes an important share of vehicles primary used for commercial purpose. We shall thus assume that the remaining share is negligible, so that vehicles with commercial registration are supposed to correspond to employer-owned vehicles and those with at least a 75 percent commercial use rate. By excluding the latter from the NaPVUS, we thus obtain comparable samples.

Note that while the previous adjustments consisted of adding quantities to the estimates derived from the survey with the narrowest sampling field in order to make them comparable to the other ones, the latest adjustment operates in the exact opposite direction. The rationale for removing primary commercial-use vehicles from the NaPVUS is that resulting estimates will be easily interpreted as applying to (primarily) private-use vehicles. Misleading conclusions could be drawn otherwise, because the characteristics of a vehicle and its users can vary considerably depending on its primary use.

For example, the number of primarily commercial-use vehicles can be inferred from the difference between the total number of registrations and the estimated number of primarily private-use vehicles so obtained. If commercial-use vehicles were part of the survey-based estimates on the basis that any private use (however limited) is made out of them, the commercial-use portion of the total number of vehicles would appear lower than it should be because some primarily commercial-use vehicles would have been wrongly assigned to the private vehicle type.

Chapter 4

Estimation and Prediction Methods

This chapter proposes a method for dealing with the problem of incomplete information in the three time series of national interest: the private vehicle stock, the average distance they travel and their weighted fuel consumption rates. A simultaneous equations model that encompasses these variables is formulated. Sampling-based techniques incorporating a data augmentation stage are applied to fill the gap between the two surveys to provide us with aggregate data on these variables. Provided that the main purpose of this study lies in the prediction of these time series, only the estimation of an unrestricted reduced form of the model is required.

The estimation of the simultaneous equations model is first undertaken equation by equation to identify relevant predictors and to reveal the underlying dynamics. We then turn to the estimation of the system of equations as a whole in order to account for the interrelationships between the three transport aggregates. The next chapter will present the estimation and prediction results obtained by applying the Bayesian estimation methods developed in this chapter.

4.1. Econometric model

The fuel consumption identity (1) introduced in Section 2.1 of Chapter 2 underlies our prediction model. This identity factorizes total fuel consumption, C, as the product of three components:

$$C=S\left(\frac{D}{S}\right)\left(\frac{C}{D}\right)=S\overline{D}E \tag{1}$$

where S stands for the vehicles stock and D is the total distance they travel. Hence, $\overline{D}$=D/S represents the average distance travelled by each vehicle, and E=C/D is their fuel efficiency, or weighted fuel consumption rates. Note that these components correspond exactly to the variables of interest to be modelled.

For reasons given in Chapter 2, Schimek's (1996) model was selected to study the three energy components. Geared to longitudinal studies of private-vehicle use aggregated time series, the model embeds the three variables of interest in a simultaneous equations framework. In its most general form, the structural model can be written as follows:

$$\begin{aligned} S_t &= f(\overline{D}_t, E_t, x_{1t}) + \varepsilon_{1t} \\ \overline{D}_t &= g(S_t, E_t, x_{2t}) + \varepsilon_{2t} \\ E_t &= h(S_t, \overline{D}_t, x_{3t}) + \varepsilon_{3t} \end{aligned} \qquad (2)$$

where the t subscript refers to the current time period, $f(\cdot)$, $g(\cdot)$ and $h(\cdot)$ represent functions whose first two arguments correspond to the two other energy components while the third ones, the x_{it}'s, stand for row vectors of exogenous explanatory variables of dimension $1xk_i$, and the ε_{it}'s are error terms relating to the i^{th} equation in the system, for $i=1,2,3$ and $t=1,\ldots,T$.

Since the estimation of a general non-linear simultaneous equations model can be complicated, Schimek assumes a linear form for the functions $f(\cdot)$, $g(\cdot)$ and $h(\cdot)$ appearing in equation (2). We make the same assumption, because of the additional difficulty raised by the fact that the three time series to be modelled involve missing data. Under the linear assumption, the unrestricted reduced form of the structural model (2) becomes:

$$y_t' = r_t \Pi + e_t \qquad (3)$$

where $y_t' = [S_t \ \overline{D}_t \ E_t]$ and $e_t = [e_{1t} \ e_{2t} \ e_{3t}]$ are vectors of dimension 1xn, n=3 being the number of equations in the system, r_t is a row vector of dimension r containing the distinct elements in $\{x_{1t}, x_{2t}, x_{3t}\}$, and Π is the associated parameter matrix of dimension rxn.

The main purpose of the present analysis lies in the estimation of the unrestricted reduced form (3), which is sufficient to obtain predictors for the three dependent variables. Once predictions have been obtained for the three variables forming the system of equations (3), fuel consumption can be derived from equation (1) by simply multiplying these three quantities. Likewise, the total distance travelled by all private-use vehicles can be obtained by multiplying the vehicle stock by the average distance travelled by each vehicle.

If the error terms e_{1t}, e_{2t}, and e_{3t}, in the system of equations (3) were independent of one another,[1] estimating the unrestricted reduced form would reduce in estimating each equation separately. Consequently, if the dependent variables formed complete time series for the period of interest, parameter estimates could be calculated by means of independent ordinary least squares regressions. However, such is not the case, so we must use the methods described in the following sections to obtain the desired predictions.

4.2. Single equation model estimation

The basic extrapolation methodology aiming at bridging the gap between the two survey estimates relies on a Bayesian approach using Gibbs sampling with data augmentation. The Gibbs sampler, introduced in Section 2.5.2 of Chapter 2, enables the Bayesian estimation to rely on the full set of conditional posterior distributions of the parameters, instead of their joint posterior distribution. By considering missing values of the dependent variables as further parameters for estimation, the data augmentation procedure allows us to obtain predictions for the intervening period between the two surveys.

For simplicity, first assume that the three error terms in equation (3) are not correlated. This rather unrealistic assumption will be relaxed in the following section. The

[1] The only impact of the assumed independence of error terms is on the efficiency of the estimates and predictions. Specifically, if the assumption were to be violated, the resulting estimates and predictions would be less efficient, but still consistent.

simpler estimation framework considered here provides benchmarks to which results derived from more convenient, but more complex, estimation methods can be compared and contrasted. Under the assumption of contemporaneous independence, a simple linear regression (LIN) model can be formulated for each equation separately:

$$y = X\beta + u \tag{4}$$

where β is a parameter vector of dimension kx1, X is the corresponding explanatory variable matrix of dimension Txk, u is a Tx1 vector of error terms, and the dependent variable comprises three components: $y=(y_F',y_M',y_N')'$. The (T_1-1)x1 vector y_F, and the $(T-T_2)$x1 vector y_N correspond, respectively, to observations on the dependent variable from the FCS and the NaPVUS, where $[T_1,T_2]$ is the interval between the two surveys. The (T_2-T_1+1)x1 vector y_M represents observations for which predictions are required for the intervening period. The resulting augmented sample of T observations including predictions is said to be complete, as opposed to the sample of observable data.

4.2.1. Simple linear regressions

Suppose, to begin with, that the error terms in model (4) are independently and normally distributed with zero mean and variance σ^2: $u \sim N(0, \sigma^2 I_T)$, where I_T stands for the identity matrix of rank T. The (T_2-T_1+k+2)x1 parameter vector for which estimations are required is composed of elements of β, σ^2 and y_M: $\theta=(\beta',\sigma^2,y_M')'$. In the present framework, natural conjugate priors is a standard choice. Specifically, a combination of a Normal distribution for the parameter vector β and an inverted Gamma distribution for the variance parameter σ^2 is selected. Conditional on values of β and σ^2, the posterior distribution of y_M is fully determined by model (4).

If the prior on β is $p(\beta)=N(b,A^{-1})$, then its conditional posterior distribution,[2] which is also normal, is given by:

$$p(\beta \mid \sigma^2, X, y) = N(\bar{\beta}, \bar{A}^{-1}) \tag{5}$$

where $\bar{A} = X'X + A$ and $\bar{\beta} = \bar{A}^{-1}(X'y + Ab)$. Moreover, if $1/\sigma^2$ is Gamma distributed, that is $p(1/\sigma^2)=G(c/2,d/2)$,[3] then its conditional posterior distribution has the following form:

$$p(1/\sigma^2 \mid \beta, X, y) = G\left(\frac{c+T}{2}, \frac{d+u'u}{2}\right) \tag{6}$$

where $u'u = (y - X\beta)'(y - X\beta)$ denotes the sum of squared residuals in model (4).

Combined with the Gibbs sampler, the data augmentation procedure generates values for the latent variables as well as values for the other model parameters from their complete set of conditional posterior distributions given the model specification, the priors, and the current values of the other parameters. Precisely, starting from arbitrary values β° and $(\sigma^2)^{\circ}$ for the model parameters β and σ^2, respectively, the Gibbs sampler generates at the i^{th} iteration:

i) (data augmentation stage) values for the latent variables from:

$$y^i_{Mt} = x_{Mt}\beta^{i-1} + u^{i-1}_{Mt} \quad \text{where } u^{i-1}_{Mt} \sim NID\left(0, (\sigma^2)^{i-1}\right) \tag{7}$$

and β^{i-1} and $(\sigma^2)^{i-1}$ correspond, respectively, to the values of β and σ^2 computed during the preceding iteration;

ii) given $y^i = (y_F', y^i_M{}', y_N')'$, an updated value, β^i, of β from its conditional posterior distribution:

$$\beta^i \sim p(\beta \mid (\sigma^2)^{i-1}, X, y^i) = N(\bar{\beta}^i, \bar{A}^{-1}) \tag{8}$$

where $\bar{A} = (X'X + A)$ and $\bar{\beta}^i = \bar{A}^{-1}(X'y^i + Ab)$;

[2] See Zellner (1971) for a derivation of the conditional posterior distributions.

[3] Note that the notation of parameters of the Gamma distribution may differ in the literature. The one we make use of here is such that if Z~G(c,d), then the random variable Z has mean E[Z]=c/d and variance V(Z)=c/d^2.

iii) given β^i and y^i, an updated value $(\sigma^2)^i$ of σ^2 from the conditional posterior distribution of $1/\sigma^2$:

$$\frac{1}{(\sigma^2)^i} \sim p\left(1/\sigma^2 \mid \beta^i, X, y^i\right) = G\left(\frac{c+T}{2}, \frac{d+u^{i\prime}u^i}{2}\right), \text{ where } u^i=y^i-X\beta^i. \qquad (9)$$

Postulated prior parameter values are: b=1, $A=I_k$, c=0.5 and d=2. OLS estimates based on the reduced sample without missing information (obtained by simply stacking the observations from both surveys) provide good starting values for the model parameters. The Bayesian estimates, $\hat{\beta}$ and $\hat{\sigma}^2$, are obtained from the means of the values generated after the burn-in period has elapsed. The predictions for the missing data, $\hat{y}_M$, are calculated in a similar fashion as the averages of the y_M values generated in each iteration following the burn-in period. Since the predictions of a linear regression model simply represent a linear transformation of the β parameter vector, namely $r(\theta)=r(\beta)=X\beta$, they could also be easily obtained from:

$$\hat{y}=E[r(\beta) \mid X]=XE[\beta \mid X]= X\hat{\beta} \qquad (10)$$

4.2.2. Adding lagged dependent variables as regressors

Predictions resulting from the application of the Gibbs sampling method with data augmentation to the simple linear regression model provide a yardstick for those generated by more general models and estimation methods. Since we are dealing with quarterly time series, it is crucial to allow for the dynamic dependencies explicitly in the model. The first attempt in that direction consists in including lagged dependent variables as a regressors within the simple linear regression model (4) in order to account for serial correlation in the time series.

The value of a seasonal variable in a given quarter is generally correlated with its value four periods ahead. The vehicle stock is one exception to that rule. The current vehicle stock is equal to the number of vehicles that were in operation in the preceding quarter plus the new vehicle sales minus the number of vehicles that were send to scrap in the interval. Accordingly, the vehicle stock in the preceding quarter is a more important determinant of the current number of vehicles than the stock prevailing during the same quarter last year. Had we considered the new vehicles sales instead, the number of vehicles purchased at the same season last year would have been more relevant to the explanation of the actual sales.

In order to account for serial correlation in the series for which predictions are required, we included lags of the dependent variable as additional regressors within the simple linear regression model (4). Models involving p lagged dependent variables (GLAG(p)) may be represented in the following way:

$$y_t = \alpha(L)y_t + x_t\beta + u_t = \sum_{s=1}^{p}\alpha_s y_{t-s} + x_t\beta + u_t \quad \text{where } u_t \sim NID\left(0,\sigma^2\right) \qquad (11)$$

and $\alpha(L) = \sum_{s=1}^{p}\alpha_s L^s$ is a p-order polynomial in the lag operator L which, whenever applied to the current period variable, y_t, results in: $L^s y_t = y_{t-s}$. We also consider a restricted version of this model (PLAG(4)) that incorporates a single 4-period lagged dependent variable.

The use of lagged dependent variables in the presence of incomplete information raises a new challenge from a methodological viewpoint. Now, missing values do not only appear in the dependent variable vector, but also in the matrix of explanatory variables. We propose a generalization of the MCMC method introduced in the preceding section which solves this additional difficulty. In order to do so, we make the most of the iterative nature, the convergence and non-simultaneity properties of the Gibbs sampling algorithm. Since the data augmentation stage can be operated one quarter at the time, the value of y_M for the tth observation generated at the ith iteration of the process, y^i_{Mt}, may be used as a regressor s periods ahead. Based on this principle, we may form the extended row vector of explanatory variables: $z^{i}_{Mt+s} = \left[y^{i}_{Mt} \quad y^{i}_{Mt+1} \cdots y^{i}_{Mt+s-1} \; x_{t+s}\right]$ to be used in the calculation of y^{i}_{Mt+s}.

The value of the explanatory variable y_{t-s} is known for the first p observations immediately following the end of the first survey period, although that of the current dependent variable y_{Mt} must be computed. Once the values of y_{Mt} are determined, they serve as the independent variables in the regressions used to compute the values of y_{Mt+s} for the next s=1,...,p observations, and so on. Note that this procedure satisfies the general idea behind the Gibbs sampling method. First, it uses the conditional posterior distribution of the dependent variable to complete the series. Second, it incorporates, in each step, all the available and most up-to-date information regarding the parameter vector for which estimates are required, including the missing values of the dependent variable. We do condition on the first p observations of the sample, that is these observations solely contribute to initialize the lagged dependent variables used as regressors p periods ahead. The initial observations do not enter the estimation process otherwise.

The Gibbs sampling algorithm used to estimate model (11) starts from arbitrary values α^0, β^0 and $(\sigma^0)^2$, for the model parameters $\alpha=(\alpha_1,\ldots,\alpha_p)'$, β and σ^2, respectively, and generates at the ith iteration:

i) values for the error terms as $u^{i-1}_t \sim NID(0,(\sigma^2)^{i-1})$ with the help of which we can compute values for the missing dependent variables from:

$$y^{i}_{Mt} = \sum_{s=1}^{p} \alpha^{i-1}_s y^{*}_{t-s} + x_t\beta^{i-1} + u^{i-1}_t \qquad (12)$$

where $y^{*}_{t-s} = \begin{cases} y_{t-s} & \text{for } t = T_1,\ldots,T_1+(p-1) \\ y^{i}_{Mt-s} & \text{for } t = T_1+p,\ldots,T_2 \end{cases}$

while $\alpha^{i-1}, \beta^{i-1}$ and $(\sigma^2)^{i-1}$ correspond, respectively, to the values of the parameters α, β et σ^2 computed during the preceding iteration. Note that the last p variables generated in this way will serve as explanatory variables for observations $T_2+1,\ldots,T_2+p$;

ii) given the completed dependent variable vector, y^i, the completed explanatory variable matrix, Z^i, and $(\sigma^2)^{i-1}$, updated values $\delta^i=(\alpha^{i\prime},\beta^{i\prime})'$ for the parameter vector $\delta=(\alpha',\beta')'$ from its conditional posterior distribution:

$$\delta^i \sim p\left(\delta \mid (\sigma^2)^{i-1}, Z^i, y^i\right) = N\left(\bar{\delta}^i, \left(\bar{A}^i\right)^{-1}\right) \tag{13}$$

where $\bar{A}^i=\left(Z^{i\prime}Z^i+A\right)$, $\bar{\delta}^i=\left(\bar{\alpha}^{i\prime},\bar{\beta}^{i\prime}\right)=\left(\bar{A}^i\right)^{-1}\left(Z^{i\prime}y^i+Ab\right)$, and it is assumed that the marginal prior on δ is: $p(\delta)=N(b,A^{-1})$;

iii) given updated values δ^i, Z^i and y^i, an updated value, $(\sigma^2)^i$, of σ^2 from the conditional posterior distribution of $1/\sigma^2$:

$$\frac{1}{\left(\sigma^2\right)^i} \sim p\left(1/\sigma^2 \mid \delta^i, Z^i, y^i\right) = G\left(\frac{c+T}{2}, \frac{d+\sum_{t=1}^{T}\left(u_t^i\right)^2}{2}\right) \tag{14}$$

where $u_t^i = y_t^i - z_t^i\delta^i = y_t^i - \sum_{s=1}^{p}\alpha_s^i y_{t-s}^i - x_t\beta^i$ for $t=1,\ldots,T$.

Initial values of the parameters can be set to the Bayesian estimates obtained for the simple linear model of the preceding section with $\alpha^0=0$. The same values as in the above section are also postulated for the prior parameters.

4.2.3. Autocorrelation treatment

Model (4) does not only involve time series in the form of the dependent variable for which predictions are required, but also in the explanatory variable matrix. The hypothesis of independent error terms is therefore likely to be violated. In a second attempt to control for possible inter-temporal dependencies in the time series, we allow for the model error terms to be correlated in time. Since we are dealing with quarterly data, we consider a fourth-order autocorrelation structure (AR(4)) of the form:

$$\phi(L)e_t = e_t - \phi_1 e_{t-1} - \phi_2 e_{t-2} - \phi_3 e_{t-3} - \phi_4 e_{t-4} = u_t \quad \text{where } u_t \sim NID(0,\sigma^2) \tag{15}$$

and $\phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \phi_3 L^3 - \phi_4 L^4$ is a polynomial in the lag operator L.

As mentioned in Section 2.5.3 of Chapter 2, we follow Chib and Greenberg's (1994) treatment of autocorrelation, but generalize the method to include a data augmentation stage, in order to obtain predicted values of the dependent variable for the transient period, between the two surveys. The authors address the more general issue of estimating the posterior distributions of the parameters for a model with an autoregressive moving average structure for the error terms (ARMA(p,q)). However, they also suggest a simplified algorithm for dealing with models involving no moving average component, which is convenient for our needs. Building on the previous paper by Chib (1993), only concerned with autoregressive models, the proposed MCMC technique does not impose conditioning on the first p=4 observations, but rather assumes that they are drawn from the stationary distribution:

$$Y_1 = (y_1,\ldots,y_p)' \sim p(Y_1 \mid \beta,\sigma^2,\phi) = N(X_1\beta,\sigma^2\Sigma) \qquad (16)$$

where X_1 contains the first p rows of X, the Σ matrix satisfies $\Sigma = \Omega\Sigma\Omega' + \iota(p)\iota(p)'$ where $\iota(p) = (1,0,\ldots,0)'$ is the px1 unit vector, and Ω is a matrix defined as follows:

$$\Omega = \begin{bmatrix} \phi'_{-p} & \phi_p \\ I_{p-1} & 0 \end{bmatrix} \text{ with } \phi_{-p} = (\phi_1,\ldots,\phi_{p-1})'.$$

Partition the model components as follows: $y=(Y_1',Y_2')'$, $X = [X_1\ X_2]$ and $u=(u_1',u_2')'$, where Y_1 and u_1 are px1 vectors corresponding to the first p observations, Y_2 and u_2 are (T-p)x1 vectors involving the remaining observations, while X_1 and X_2 are the corresponding explanatory variable matrices of dimensions pxk and (T-p)xk, respectively. For simplicity, we assume that X_1 and X_2 do not involve any lagged dependent variables, as in Chib and Greenberg. Then, consider the transformed variables:

$$\begin{cases} \tilde{Y}_1 = Q^{-1}Y_1, & \tilde{X}_1 = Q^{-1}X_1, \quad u_{1t} = \tilde{y}_{1t} - \tilde{x}_{1t}\beta \text{ for } t = 1,\ldots p \\ \tilde{y}_{2t} = \phi(L)y_t, & \tilde{x}_{2t} = \phi(L)x_t, \quad u_{2t} = \tilde{y}_{2t} - \tilde{x}_{2t}\beta \text{ for } t = p+1,\ldots,T \end{cases} \qquad (17)$$

where Q stands for the Cholesky matrix of Σ, defined such that Σ=QQ'. In stacked form, let $\tilde{y} = (\tilde{Y}_1', \tilde{Y}_2')'$ and $\tilde{X} = [\tilde{X}_1 \; \tilde{X}_2]$. Then, the transformed model may be written as:

$$\tilde{y} = \tilde{X}\beta + u \quad \text{where } u \sim N\left(0, \sigma^2 I_T\right) \tag{18}$$

The following priors are postulated. β and σ^2 are assumed to follow the usual Normal-inverted-Gamma priors: p(β)=N(b,A^{-1}) and p(1/σ^2)=G(c/2,d/2), while ϕ follows a truncated normal distribution: $p(\phi) = N(\varphi, \Phi^{-1}) S_\phi$, restricted to the region, S_ϕ, of stationary values of ϕ, that is S_ϕ contains parameter vectors ϕ insuring that all roots of the ϕ(L) polynomial lie outside of the unit circle. Lets further define the (T-p)x1 vector of autocorrelated error terms $\varepsilon = (e_{p+1}, \ldots, e_T)'$ and the corresponding (T-p)xp matrix, E, of lagged errors whose tth row is given by $[e_{t-1} \; \ldots \; e_{t-p}]$. Under the preceding hypotheses and assuming, for the moment, that the y vector is complete, the conditional posterior distributions of the model parameters are:

$$p(\beta \mid \sigma^2, \phi, y, X) = N(\overline{\beta}, \overline{A}^{-1}) \tag{19}$$

$$p(1/\sigma^2 \mid \beta, \phi, y, X) = G\left(\frac{c+T}{2}, \frac{d+u'u}{2}\right) \tag{20}$$

$$p(\phi \mid \beta, \sigma^2, y, X) \propto \Psi(\phi)\, N\left(\overline{\varphi}, \overline{\Phi}^{-1}\right) S_\phi \tag{21}$$

where $\overline{A} = \sigma^{-2}\, \tilde{X}'\tilde{X} + A$, $\quad \overline{\beta} = \overline{A}^{-1}\left(\sigma^{-2}\tilde{X}'\tilde{y} + Ab\right)$,

$\overline{\Phi} = \sigma^{-2}E'E + \Phi$, $\quad \overline{\varphi} = \overline{\Phi}^{-1}\left(\sigma^{-2}E'\varepsilon + \Phi\varphi\right)$,

and $\quad \Psi(\phi) = |\Sigma|^{-1/2} \exp\left\{\frac{-1}{2\sigma^2} u_1' u_1\right\}$.

Distributions (19) and (20) are standard, but that is not the case for distribution (21) which deserves more attention. In order to simulate random variates from the conditional posterior distribution of ϕ, we will appeal to another MCMC technique from

Chapter 2: the Metropolis-Hastings (M-H) algorithm. This powerful tool allows one to generate random variates, z, from a multidimensional density function, p(z)=cr(z), where c, the normalizing constant, might be unknown, by drawing random variates from another density function, called the candidate-generating density, suitably chosen for convenience and ease of implementation. Given the current random variates, z^i, the M-H algorithm draws a candidate, z^*, from the candidate-generating density $q(z^i,z^*)$. The process then sets the next value of z to $z^{i+1}=z^*$ provided a certain condition is satisfied; otherwise, it returns $z^{i+1}=z^i$.

The optimal condition would be that the candidate-generating density satisfies the reversibility condition defined as:

$$p(x)q(x,y) = p(y)q(y,x)$$

for all admissible values of x and y, in which case p() would be an invariant density.[4] If the candidate-generating density respected the reversibility condition, then the algorithm would generate convergent values from p() at each iteration. However, in general, this is not the case and for some values of x and y, we have:

$$p(x)q(x,y) > p(y)q(y,x)$$

which means that the process moves from x to y too often and from y to x too rarely.

In order to correct the situation, the candidate is subjected to a further randomization by introducing a probability, $m(x,y) < 1$, that the move is made. The probability of move is chosen in such a way that the transition probability, $q(x,y)m(x,y)$, meets the reversibility condition when $m(x,y) = 1$. In this case,

$$p(x)q(x,y)m(x,y) = p(y)q(y,x)m(y,x) = p(y)q(y,x).$$

[4] For more details on these conditions and their relationship to the M-H algorithm, see Tierney (1993) or Chib and Greenberg (1995).

Therefore, by definition,

$$m(x,y) = \frac{p(y)q(y,x)}{p(x)q(x,y)} \tag{22}$$

In other words, the candidate, z*, is accepted with probability

$$m\left(z^i,z^*\right) = \min\left\{\frac{p(z^*)q\left(z^*,z^i\right)}{p\left(z^i\right)q\left(z^i,z^*\right)},1\right\}.$$

Under mild regularity conditions, the draws from the M-H algorithm converge almost surely to the invariant distribution p(z) after the burn-in period.[5] Note that knowledge about the normalizing constant, c, is not necessary because it appears in both the numerator and the denominator of the probability of move, as defined in (22). Furthermore, provided that the candidate generating density is symmetrical, q(y,x)=q(x,y) for all values of x and y, the probability of move reduces to $m(x,y) = \min\{p(y)/p(x),1\}$. Similarly, whenever p(z) may be written as $p(z) \propto \Xi(z)h(z)$, where h(z) is a density that can be sampled from, and $\Xi(z)$ is uniformly bounded, and when the candidate-generating density is chosen to be q(x,y)=h(y), the probability of move reduces to $m(x,y) = \min\{\Xi(y)/\Xi(x),1\}$.

As noted in Section 2.5.2 of Chapter 2, a M-H step might be combined with the Gibbs sampling process to generate some of the conditional posterior densities. This is exactly what is required here to generate random variates from the non-standard conditional posterior distribution of ϕ. A natural and efficient choice for the candidate-generating density[6] is inspired by the form of (21), namely $N\left(\phi \mid \bar{\phi},\bar{\Phi}^{-1}\right)S_\phi$. Since $\Psi(\phi)$ is proportional to a normal distribution, it can be uniformly bounded and so, as discussed in

[5] See Smith and Roberts (1993) for a more detailed discussion of the convergence issues.

[6] Refer to Chib and Greenberg (1994, 1995) for a justification.

the preceding paragraph, the probability of move reduces to $m(\phi^i,\phi^*)=\min\{\Psi(\phi^*)/\Psi(\phi^i)1\}$.

In summary, the Gibbs sampling algorithm, augmented with a M-H step proposed by Chib and Greenberg (1995) to estimate a model with autocorrelation, and further generalized to incorporate a data augmentation step, proceeds as follows. Starting from arbitrary values β^0, ϕ^0 and $(\sigma^0)^2$, for the model parameters β, ϕ and σ^2, respectively, at the ith iteration the algorithm:

i) computes values for the last four autocorrelated error terms immediately preceding the end of the FCS sampling period as follows:

$e_t^{i-1} = y_t - X_t\beta^{i-1}$ for $t = T_1 - 4,\ldots,T_1 - 1$,

generates values for the independent error terms within the intermediary period between the two surveys from:

$u_t^{i-1} \sim \mathrm{NID}(0,(\sigma^2)^{i-1})$ for $t = T_1,\ldots,T_2$,

and, from these, computes values for the autocorrelated error terms within the same intermediary period as:

$e_t^{i-1} = \phi_1^{i-1}e_{t-1}^{i-1} + \phi_2^{i-1}e_{t-2}^{i-1} + \phi_3^{i-1}e_{t-3}^{i-1} + \phi_4^{i-1}e_{t-4}^{i-1} + u_t^{i-1}$ for $t = T_1,\ldots,T_2$.

Using these values, compute values for the missing dependent variables as:

$$y_{Mt}^i = X_t\beta^{i-1} + e_t^{i-1} \quad \text{for } t = T_1,\ldots,T_2, \qquad (23)$$

where β^{i-1}, ϕ^{i-1} and $(\sigma^2)^{i-1}$ correspond, respectively, to the values of the parameters α, β and σ^2 computed during the preceding iteration;

ii) given the completed dependent variable vector, y^i, computes transformed variables, $\tilde{y}^i$ and $\tilde{X}^i$, in expression (17) using $\phi=\phi^{i-1}$, and generates updated values for the β parameter vector, β^i, from its conditional posterior distribution:

$$\beta^i \sim p\left(\beta \mid \phi^{i-1},\left(\sigma^2\right)^{i-1},y^i,X\right) = N\left(\overline{\beta}^i,\left(\overline{A}^i\right)^{-1}\right) \qquad (24)$$

where $\overline{A}^i = \left(\sigma^{-2}\right)^{i-1}\tilde{X}^{i\prime}\tilde{X}^i + A$ and $\overline{\beta}^i = \left(\overline{A}^i\right)^{-1}\left[\left(\sigma^{-2}\right)^{i-1}\tilde{X}^{i\prime}\tilde{y}^i + Ab\right]$;

iii) given updated values β^i and y^i, generates an updated value, $(\sigma^2)^i$, of σ^2 from the conditional posterior distribution of $1/\sigma^2$:

$$\frac{1}{(\sigma^2)^i} \sim p\left(1/\sigma^2 \mid \beta^i, \phi^{i-1}, y^i, X\right) = G\left(\frac{c+T}{2}, \frac{d+\sum_{t=1}^{T}\left(u_t^i\right)^2}{2}\right) \tag{25}$$

where $u_t^i = \tilde{y}_t^i - \tilde{X}_t^i\beta^i$ for t=1,…,T;

iv) given updated values β^i, $(\sigma^2)^i$ and y^i, generates an updated value, ϕ^i, of ϕ from the following M-H step:

- first generates a value ϕ^* from the candidate-generating density:

  $$\phi^* \sim N\left(\overline{\varphi}^i, \left(\overline{\Phi}^i\right)^{-1}\right) S_\phi \tag{26}$$

  where $\overline{\Phi}^i = \left(\sigma^{-2}\right)^i E^{i\prime}E^i + \Phi$, $\overline{\varphi}^i = \left(\overline{\Phi}^i\right)^{-1}\left(\left(\sigma^{-2}\right)^i E^{i\prime}\varepsilon^i + \Phi\varphi\right)$ and ε^i and E^i are defined as in (30) using $e_t^i = y_t^i - X_t\beta^i$ for t=p+1,…,T;

- then draws a pseudo-random number, w, from the uniform distribution on the [0,1] interval;

- computes $m\left(\phi^{i-1}, \phi^*\right) = \min\left\{\Psi(\phi^*)/\Psi\left(\phi^{i-1}\right)1\right\}$ where $\Psi(\phi)$ is defined as:

  $$\Psi(\phi) = |\Sigma|^{-1/2}\exp\left\{\frac{-1}{2\left(\sigma^2\right)^i}\left(Y_1 - X_1\beta^i\right)\Sigma^{-1}\left(Y_1 - X_1\beta^i\right)\right\};$$

- if $w \le m\left(\phi^{i-1}, \phi^*\right)$, sets $\phi^i = \phi^*$; otherwise sets $\phi^i = \phi^{i-1}$.

The following values are assumed for the prior on the autocorrelation parameters: φ_i=0.1 for i=1,…,p and $\Phi = I_p$. The other parameters of the model keep the same prior parameter distribution values as above. OLS regressions performed on the stacked survey-based data provide starting values for the β and σ^2 parameters. In order to be perfectly consistent with OLS estimates, null starting values should be assigned to all autocorrelation parameters. But since the algorithm runs into difficulties whenever such limiting values are used as initial values, we rather start from ϕ_i=0.1 for i=1,…,p. Note that these values satisfy the stationarity conditions.

4.3. Simultaneous equations model estimation

Consider the more general problem consisting in the estimation of the system of equations (3) as a whole. In the following sections, we first estimate a static model and then relax some of the underlying assumptions to introduce some dynamics. The next section explains how we may estimate a seemingly unrelated regression (SUR) model using the Gibbs sampler. The following ones address the more general issue of Bayesian estimation of vector autoregressive (VAR) models.

4.3.1. Seemingly unrelated regressions

The unrestricted reduced form of the system (3) may be rewritten in the form of a seemingly unrelated regression (SUR) model, where each equation involves the same explanatory variable matrix. However, for reasons that will become apparent when we analyse the results of Monte Carlo experiments, we do not impose that each equation involves exactly the same set of regressors and we write the seemingly unrelated regression model as follows:

$$y_t = X_t\theta + \varepsilon_t \tag{27}$$

where $y_t = (S_t, \overline{D}_t, E_t)'$ and $\varepsilon_t = (\varepsilon_{1t}, \varepsilon_{2t}, \varepsilon_{3t})'$ are vectors of dimension nx1, θ is a kx1 vector of parameters, and the corresponding explanatory variable matrix X_t of dimension nxk is defined as:

$$X_t = \begin{bmatrix} x_{1t} & 0 & 0 \\ 0 & x_{2t} & 0 \\ 0 & 0 & x_{3t} \end{bmatrix}, \tag{28}$$

with $k=k_1+k_2+k_3$. Note that the notation may be modified in an obvious way to allow for parameter equality restrictions across equations, although we do not impose such restrictions here.

For the moment, let's assume that the X_t matrices in (27) do not involve any lagged dependent variables. Furthermore, consider the hypothesis that the error terms are distributed according to: $\varepsilon_t \sim \text{NID}(0,\Sigma)$, where Σ is an nxn positive definite matrix, for t=1,...,T. In other words, error terms are uncorrelated from one period to another, but they can be contemporaneously correlated. Let Q be the Cholesky matrix of Σ^{-1} defined in such a way that $QQ'=\Sigma^{-1}$, and consider the transformed model:

$$\tilde{y}_t = \tilde{X}_t\theta + u_t \tag{29}$$

where the transformed variables are defined in the following way: $\tilde{y}_t = Qy_t$, $\tilde{X}_t = QX_t$ and $u_t = Q\varepsilon_t \sim \text{NID}(0,I_n)$.

Under the usual natural conjugate prior that $p(\theta)=N(b,A^{-1})$ and $p(\Sigma^{-1})=W(\nu,\Lambda)$, where W denotes the Wishart distribution,[7] Zellner (1971) shows that the conditional posterior distributions for θ and Σ^{-1} are, respectively:

$$p(\theta \mid \Sigma, X, y)=N\left(\bar{\theta},\bar{A}^{-1}\right) \tag{30}$$

$$p\left(\Sigma^{-1} \mid \theta, X, y\right)= W(\bar{\nu},\bar{\Lambda}) \tag{31}$$

where $\bar{A}=\left(\sum_{t=1}^{T}\tilde{X}_t{}'\tilde{X}_t + A\right)$, $\bar{\theta}=\bar{A}^{-1}\left(\sum_{t=1}^{T}\tilde{X}_t{}'\tilde{y}_t + Ab\right)$, $\bar{\nu} = \nu + T$ and $\bar{\Lambda} = \Lambda + \sum_{t=1}^{T}\varepsilon_t\varepsilon_t{}'$.

This suggests the following Gibbs sampling scheme for estimation and predictions. Starting from arbitrary values θ° and Σ° for the model parameters θ and Σ respectively, the algorithm generates, at the i^th^ iteration:

i) values for the latent variables from:

$$y^{i}_{Mt} = X_{Mt}\theta^{i-1} + \varepsilon^{i-1}_{Mt} \quad \text{where } \varepsilon^{i-1}_{Mt} \sim \text{NID}\left(0,\Sigma^{i-1}\right) \tag{32}$$

where θ^{i-1} and Σ^{i-1} correspond respectively, to the values of θ and Σ computed during the preceding iteration;

$$y^{i}_{Mt} = X_{Mt}\theta^{i-1} + \varepsilon^{i-1}_{Mt} \quad \text{where } \varepsilon^{i-1}_{Mt} \sim \text{NID}\left(0,\Sigma^{i-1}\right)$$

[7] The Wishart distribution is the matrix extension of the Chi-square (χ^2) distribution. See Geweke (1988) for an algorithm generating Wishart pseudo-random variates.

ii) given the complete dependent variable vector y^i, an updated value θ^i of θ from its conditional posterior distribution:

$$\theta^i \sim p\left(\theta \mid \left(\sigma^2\right)^{i-1}, X, y^i\right) = N\left(\bar{\theta}^i, \left(\bar{A}^i\right)^{-1}\right) \qquad (33)$$

where $\bar{A}^i = \left(\sum_{t=1}^{T} \tilde{X}_t^{i}{}' \tilde{X}_t^i + A\right)$, $\bar{\theta}^i = \left(\bar{A}^i\right)^{-1}\left(\sum_{t=1}^{T} \tilde{X}_t^{i}{}' \tilde{y}_t^i + Ab\right)$ and the transformed variables $\tilde{X}_t^i$ and $\tilde{y}_t^i$ are formed as in (29), using the Cholesky matrix of Σ^{i-1};

iii) given θ^i and y^i, an updated value Σ^i of Σ from the conditional posterior distribution of Σ^{-1}:

$$\left(\Sigma^i\right)^{-1} \sim p\left(\Sigma^{-1} \mid \theta^i, X, y^i\right) = W\left(\nu + T, \Lambda + \sum_{t=1}^{T} \varepsilon_t^i \varepsilon_t^{i}{}'\right), \text{ where } \varepsilon_t^i = y_t^i - X_t\theta^i \qquad (34)$$

Prior distribution parameters are set to the following values: b is a unit vector of dimension kx1, ν=0, while A and Λ are identity matrices of dimensions k and n, respectively. The starting value of the Σ matrix is set to its OLS estimate from a regression on the multivariate model (27), based on the stacked sample of observations from both surveys. A corresponding starting value for θ is then drawn from its conditional posterior distribution (30) with $\Sigma=\Sigma^0$. Based on a two-by-two equations system, this sampling-based technique performs well on simulated data, as will be argued in the next chapter.

4.3.2. Pure vector autoregressions

The main drawback of the SUR model setting is that it only allows for contemporaneous correlation among the error terms of distinct equations. No other form of inter-dependencies is admitted between them. One possible way to explicitly account for such relationships would be to incorporate other dependent variables as regressors in each equation, hence resulting in a simultaneous equations model specification. However, as argued above, that kind of setting is not required for our prediction purposes. An alternative solution consists in introducing some dynamics into the system of equations, either by incorporating a number of lags of the dependent variables as explanatory variables to form a vector autoregressive (VAR) model or by allowing for the vector of error terms to be autocorrelated. Provided that enough lags are present in the VAR model specification, however, the resulting error terms should not be autocorrelated.

Under a pure vector autoregressive (PVAR(p)) model specification, the system of equations (3) becomes:

$$y_t' = \sum_{i=1}^{p} y_{t-i}'\Phi_i + d_t\Theta + e_t \tag{35}$$

where $y_t' = [S_t \; \overline{D}_t \; E_t]$ and $e_t = [e_{1t} \; e_{2t} \; e_{3t}]$ are vectors of dimension 1xn, each parameter matrix Φ_i, for i=1,...,p, has dimension nxn, d_t is a row vector of dimension d containing only deterministic variables (namely a constant and seasonal dummy variables in our case) for the moment, Θ is the corresponding parameter matrix of dimension dxn and p is the number of lags of the dependent variables included in each equation. Note that contrarily to the univariate case, here all equations include lags of the three dependent variables as regressors. Furthermore, the same number of lags of each dependent variable is assumed to appear in each equation. Consequently, every equation appearing in the system involves exactly the same set of explanatory variables.

Assume that the error terms in model (35) are distributed as $e_t \sim NID(0,\Sigma)$, for t=1,...,T, that is they can be contemporaneously correlated, but not autocorrelated. The model can be written in the more compact form:

$$y_t' = w_t\Gamma + e_t \tag{36}$$

where $w_t = [d_t \; y_{t-1}' \cdots \; y_{t-p}']$ has dimension 1xm and $\Gamma = [\Theta' \; \Phi_1' \; \ldots \; \Phi_p']'$ is the corresponding parameter matrix of dimension mxn, with m=d+np. Performing the conventional stacking into a multivariate regression model, we obtain:

$$Y = W\Gamma + E \tag{37}$$

where Y and E are matrices of dimension Txn, and W is an explanatory variable matrix of size Txm.

Since the same regressors appear in each equation of the multivariate regression model (37), stacking the columns of the W matrix yields:

$$y = (I_n \otimes W)\gamma + \varepsilon \quad \text{where } \varepsilon \sim N(0, \Sigma \otimes I_T) \tag{38}$$

with y and ε both having dimension Tnx1, and γ being mnx1.

Several kinds of priors have been considered in the literature. Kadiyala and Karlsson (1993, 1997) compare some of them, both from a theoretical and an empirical standpoint. They find that among the Minnesota prior (Litterman, 1980, 1986), the diffuse (or Jeffrey's) prior (Geisser, 1965; Tiao and Zellner, 1964), the Normal-Wishart (i.e. natural conjugate) prior, the Normal-diffuse prior (Zellner, 1971) and the extended natural conjugate (ENC) prior (Drèze and Morales, 1976), the latest two impose the less restrictive structure on the prior variance-covariance matrix of the error terms, and consequently on the parameter posterior distributions. Furthermore, their Monte Carlo experiments show that, in some cases, these two types of priors lead to better prediction performances. In counterpart, moments of the resulting posterior distributions do not have closed form solutions and hence, must be evaluated numerically by using importance-sampling or MCMC techniques.

The ENC prior involves a reparametrization of the model that complicates its interpretation. For this reason, we prefer to use the Normal-diffuse prior which leads to conditional posterior distributions similar to those obtained in the preceding section for the SUR model. Note that this choice of prior may cause the posterior distribution to be bimodal, and hence the posterior mean to have low probability, although the authors do not expect this problem to occur frequently in practice. Results displayed in the next chapter serve as confirmations that such a problem does not occur in our Bayesian estimation of VAR models based on the Normal-diffuse prior assumption.

If a Normal-diffuse prior is assumed for the model parameters γ and Σ, respectively, summarized as $\gamma \sim N(g, \Psi)$ and $\Sigma \propto |\Sigma|^{-(n+1)/2}$, then the conditional posterior distributions have the following forms:

$$p(\gamma \mid \Sigma, y, W) \sim N(\bar{\gamma}, \bar{\Psi}^{-1}) \tag{39}$$

$$p(\Sigma^{-1} \mid \Gamma, Y, W) \sim W\left(\left[\hat{E}'\hat{E} + (\Gamma - \hat{\Gamma})W'W(\Gamma - \hat{\Gamma})\right]^{-1}, T\right) \tag{40}$$

where $\bar{\Psi} = \Psi^{-1} + \Sigma^{-1} \otimes W'W$, $\bar{\gamma} = \bar{\Psi}^{-1}\left[\Psi^{-1}g + (\Sigma^{-1} \otimes W'W)\hat{\gamma}\right]$, with $\hat{\Gamma}$ and $\hat{E}$ defined, respectively, as the OLS parameter estimates and residual sum of squares from the multivariate regression model (37): $\hat{\Gamma} = (W'W)^{-1}W'Y$ and $\hat{E} = Y - W\hat{\Gamma}$, and $\hat{\gamma}$ is defined accordingly.

Prior parameters are given the same values as in Kadiyala and Karlsson.[8] Prior means on all model parameters are set to zero, apart from those on first own lags which are set to unity:[9] $E[\Theta]=0$, $E[\Phi_1]=I_n$, and $E[\Phi_j]=0$ for $j=2,\ldots,p$. Prior covariances are all null, while prior variances for a parameter associated with an explanatory variable belonging to equation j are defined as follows:

$V((\Phi_{jj})_i) = \pi_1 / i$ for parameters on own lags, j, of length i;

$V((\Phi_{jk})_i) = \pi_2\sigma_j^2 / i\sigma_k^2$ for parameters on lags of length i of variable $k \neq j$;

$V(\Theta_{jk}) = \pi_3\sigma_j^2$ for parameters on the kth exogenous variable;

[8] See Litterman for a discussion and motivations of these prior beliefs. Kadiyala and Karlsson employ the same parameter values as does Litterman, except that variances of parameters on lagged dependent variables of length q decrease at a rate 1/q instead of at a faster rate of $1/q^2$.

[9] Note that the random walk prior of Litterman (1980) does not impair on the stationarity of posterior distributions of parameters on first own lags, but does not rule out explosive process either.

where σ_j^2 is the residual sum of squares of a p-lag univariate autoregression for the dependent variable j. We adopt the commonly used range of values for the hyper-parameters: π_1=0.05, π_2=0.005 and π_3=10^5.

Note that, as in the corresponding univariate case of the GLAG(p) model introduced in Section 4.2.2, the regression matrix involves missing variables referring to lags of dependent variables, for which no information is available. Moreover, the inclusion of these variables as regressors prevents the use, in the estimation process, of as many initial observations as the number of lags involved in such a model. In other words, the first p observations only serve for initialization purpose. Using the above conditional posterior distributions, the following Gibbs sampling algorithm with data augmentation can be implemented. Starting from arbitrary values of the model parameter vector, γ^o (with associated matrix Γ^0), and the variance-covariance matrix, Σ^o, the algorithm generates, at the ith iteration:

i) values for the error terms as $e_t^{i-1} \sim \mathrm{NID}(0,\Sigma^{i-1})$ with the help of which we can compute values for the missing dependent variables from:

$$y_{Mt}^{i}{}' = \begin{cases} w_t\Gamma^{i-1} + e_t^{i-1} & \text{for } t = T_1 \\ w_{Mt}^{i}\Gamma^{i-1} + e_t^{i-1} & \text{for } t = T_1+1,\ldots,T_2 \end{cases} \tag{41}$$

where Γ^{i-1} and Σ^{i-1} correspond, respectively, to the values of Γ and Σ computed during the preceding iteration, and w_{Mt}^{i} is defined similarly to w_t with the missing values of the lagged dependent variables replaced by the values of these variables computed for past observations. For instance, $w_{MT_1+1}^{i} = \left[d_{T_1+1}\ y_{MT_1}^{i\,'}\ y_{T_1-1}^{'} \cdots\ y_{T_1-p}^{'}\right]$ and

$$w_{MT_1+p}^{i} = \left[d_{T_1+p}\ y_{MT_1+p-1}^{i'}\ y_{MT_1+p-2}^{i'} \cdots y_{MT_1}^{i'}\right].$$

Note that the last p variables generated in this way will serve as explanatory variables for observations T_2+1,…,T_2+p;

ii) given the completed matrices Y^i and W^i created with values computed in the preceding step, OLS estimates $\hat{\Gamma}^i = (\mathrm{W}^{i\prime}\mathrm{W}^i)^{-1}\mathrm{W}^{i\prime}\mathrm{Y}^i$ (with associated parameter vector $\hat{\gamma}^i$) and $\hat{\mathrm{E}}^i = \mathrm{Y}^i - \mathrm{W}^i\hat{\Gamma}^i$ can be computed. An updated value Σ^i of Σ is then obtained from the conditional posterior distribution of Σ^{-1}:

$$\left(\Sigma^i\right)^{-1} \sim p\left(\Sigma^{-1} \mid \Gamma^{i-1}, \mathrm{Y}^i, \mathrm{W}^i\right) \sim \mathrm{W}\left(\left[\hat{\mathrm{E}}^{i\prime}\hat{\mathrm{E}}^i + \left(\Gamma^{i-1} - \hat{\Gamma}^i\right)\mathrm{W}^{i\prime}\mathrm{W}^i\left(\Gamma^{i-1} - \hat{\Gamma}^i\right)\right]^{-1}, \mathrm{T}\right) \quad (42);$$

iii) given y^i, W^i, $\hat{\gamma}^i$ and Σ^i as computed above, an updated value γ^i of γ from its conditional posterior distribution:

$$\gamma^i \sim p\left(\gamma \mid \Sigma^i, y^i, \mathrm{W}^i\right) = \mathrm{N}\left(\bar{\gamma}^i, \left(\overline{\Psi}^i\right)^{-1}\right) \qquad (43)$$

where $\overline{\Psi}^i = \Psi^{-1} + \left(\Sigma^i\right)^{-1} \otimes \mathrm{W}^{i\prime}\mathrm{W}^i$

and $\bar{\gamma}^i = \left(\overline{\Psi}^i\right)^{-1}\left[\Psi^{-1}g + \left(\left(\Sigma^i\right)^{-1} \otimes \mathrm{W}^{i\prime}\mathrm{W}^i\right)\hat{\gamma}^i\right]$.

Similarly to the SUR model treatment, starting values for the parameters of the VAR model are determined as follows. Firstly, the starting value of the error variance-covariance matrix, Σ^o, is computed as the OLS estimate of Σ from the multivariate regression model (37), based on the incomplete sample, without concern of the observations corresponding to the unobserved values of the dependent variable or its lags. Secondly, the starting value of the γ parameter vector is drawn from its conditional posterior distribution (39) using the OLS estimates computed in the preceding step.

4.3.3. Generalized vector autoregressions

In this section, we combine the tools developed in the two preceding sections in order to produce an estimation process for a general VAR model of the form (35) where the independent variable row vector may contain stochastic variables, as well as deterministic ones. The resulting Gibbs sampling algorithm is a generalization of what Kadiyala and Karlsson proposed. The method is similar to that described in the Section 4.3.2, except that the Kronecker representation of the system in form (38) is no longer valid, because the regression matrix is not necessarily the same across equations. More precisely, the explanatory variable matrix is defined based upon the corresponding line of the X_t matrix in (28) which contains zero elements in positions where the variable is not included in the given equation. Nevertheless, the same number of lags is still assumed for all dependent variables appearing in the system. The notation can be modified in an obvious way to impose equality restrictions on parameters across equations.

We consider two different representations of the generalized vector autoregressive (GVAR(p)) model: the long and short format representations. The former corresponds to a mix of the SUR representation (27) and the PVAR(p) formulation (36) and can be written as:

$$y_t = z_t\delta + \mu_t \quad \text{where } \mu_t \sim \text{NID}(0,\Sigma) \tag{44}$$

and $y_t = (S_t, \overline{D}_t, E_t)'$ and $\mu_t = (\mu_{1t}, \mu_{2t}, \mu_{3t})'$ are vectors of dimension nx1, while δ is a qx1 vector of parameters (with $q=k_1+k_2+k_3+np$) associated with the explanatory variable matrix of dimension nxq defined as follows:

$$z_t = \begin{bmatrix} x_{1t} & y'_{t-1} & \cdots & y'_{t-p} & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & x_{2t} & y'_{t-1} & \cdots & y'_{t-p} & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & x_{3t} & y'_{t-1} & \cdots & y'_{t-p} \end{bmatrix} \tag{45}$$

We obtain the equivalent of (38), which does not involve a Kronecker product, however, by stacking the columns of the matrices appearing in (45):

$$y = Z\delta + \mu \quad \text{where } \mu \sim \text{N}(0, \Sigma \otimes I_T) \tag{46}$$

and y and μ have dimension Tnx1 while Z is Tnxq.

In the short format representation corresponding to (36), the model is expressed as:

$$y_t' = s_t\Delta + u_t \tag{47}$$

where $s_t = [r_t \; y'_{t-1} \ldots y'_{t-p}]$, with the r-dimensional r_t row vector containing all distinct elements appearing in $\{x_{1t}, x_{2t}, x_{3t}\}$, as before. The parameter matrix Δ, of dimension sxn, is defined accordingly, with zero elements, $\Delta_{ij} = 0$, appearing in positions where the corresponding explanatory variable, r_{it}, does not enter in equation j, and s=r+p. Performing the conventional stacking of the observations in (47), we obtain a multivariate regression model of the same form as (37):

$$Y = S\Delta + U \tag{48}$$

where Y and U are matrices of dimension Txn, and S is an explanatory variable matrix of size Txs.

Using the above notation, and making the most of what has been learned in the two preceding sections, the following Gibbs sampling algorithm with data augmentation has been developed to estimate the GVAR model. Starting from arbitrary values of the parameter vector, δ^o (with associated matrix Δ^0), and the variance-covariance matrix, Σ^o, the algorithm produces, at the i^th^ iteration:

i) values for the error terms as $\mu_t^{i-1} \sim NID(0, \Sigma^{i-1})$ with the help of which we can compute values for the missing dependent variables from:

$$y_{Mt}^i = \begin{cases} z_t \delta^{i-1} + \mu_t^{i-1} & \text{for } t = T_1 \\ z_{Mt}^i \delta^{i-1} + \mu_t^{i-1} & \text{for } t = T_1 + 1, \ldots, T_2 \end{cases} \tag{49}$$

where δ^{i-1} and Σ^{i-1} correspond, respectively, to the values of δ and Σ computed during the preceding iteration and z_{Mt}^i is defined similarly to z_t, with the missing values of the lagged dependent variables replaced by the values of these variables computed for past observations. Note that the last p variables generated in this way will serve as explanatory variables for observations $T_2+1, \ldots, T_2+p$;

ii) given the completed dependent variable vector y^i and the corresponding explanatory variable matrices Z^i and S^i created with values computed in the preceding step, computes the OLS estimates $\hat{\delta}^i = (Z^{i\prime}Z^i)^{-1}Z^{i\prime}y^i$ and $\hat{\mu}^i = y^i - Z^i\hat{\delta}^i$, and forms the corresponding restricted estimates matrix $\hat{\Delta}^i$ with the help of which we can generate an updated value Σ^i of Σ from the conditional posterior distribution of Σ^{-1}:

$$\left(\Sigma^i\right)^{-1} \sim p\left(\Sigma^{-1} \mid \Delta^{i-1}, y^i, S^i\right) = W\left(\left[\sum_{t=1}^{T}\hat{\mu}_t^i\hat{\mu}_t^{i\prime} + \left(\Delta^{i-1} - \hat{\Delta}^i\right)S^{i\prime}S^i\left(\Delta^{i-1} - \hat{\Delta}^i\right)\right]^{-1}, T\right) \qquad (50)$$

where Δ^{i-1} stands for the value of Δ calculated during the preceding iteration;

iii) given y^i, Z^i, and Σ^i as computed above, an updated value δ^i of δ from its conditional posterior distribution:

$$\delta^i \sim p\left(\delta \mid \Sigma^i, y^i, Z^i\right) = N\left(\bar{\delta}^i, \left(\bar{\Psi}^i\right)^{-1}\right) \qquad (51)$$

where $\bar{\Psi}^i = \Psi^{-1} + \left(\sum_{t=1}^{T}\tilde{z}_t^{i\prime}\tilde{z}_t^i\right)$, $\bar{\delta}^i = \left(\bar{\Psi}^i\right)^{-1}\left[\Psi^{-1}g + \left(\sum_{t=1}^{T}\tilde{z}_t^{i\prime}\tilde{y}_t^i\right)\right]$ and the transformed variables are defined in the following way: $\tilde{y}_t^i = Q^i y_t^i$ and $\tilde{z}_t^i = Q^i z_t^i$ with Q^i being the Cholesky matrix of $(\Sigma^i)^{-1}$.

Additional model parameters are assumed to follow the same prior distributions as those corresponding to deterministic explanatory variables. Starting values for the model parameters are determined in the same way as in the preceding section.

Chapter 5

Applications

This chapter presents the results obtained from the estimation of the simultaneous equations model described in Chapter 4. First, we check the reliability and accuracy of the Bayesian estimation methods developed in the preceding chapter by means of Monte Carlo experiments based on simulated data. Second, we apply these sampling-based techniques to produce complete time series for the three transport aggregates: the private vehicle stock, the average distance they travel and their weighted fuel consumption rates. The model estimation is undertaken separately for cars, and for light trucks and vans.

This chapter presents the final results. The one following will question their validity, reliability and robustness. We start by describing the model specifications selected based on a series of diagnostic tests that will be performed in the next chapter. Alternative formulations are also provided. Recall that the model specification must rely on explanatory variables from external sources, since the survey variables would involve missing information for the intermediary period, between the surveys. We then tabulate the Bayesian estimates obtained for model types and specifications that will be shown to yield the best prediction results in the next chapter. Finally, tables and graphs of the resulting predictions are presented and discussed.

5.1. Monte Carlo experiments

In order to test the performance of each sampling-based technique described in Chapter 4, we conduct Monte Carlo experiments based on simulated data. We first check the methods applying to single equation estimation and then turn to the estimation of systems of equations. The main advantage of the simulations is that the model features, variables and parameter values are known to the investigator. Consequently, the accuracy of the Bayesian estimates can be evaluated. Furthermore, by varying the bounds of the prediction interval, an investigation of the flexibility and stability of the diverse sampling-based techniques can be undertaken.

Recall that missing information represents nearly 34% of the cars sub-sample, and 45% of the light trucks and vans sub-sample. Therefore, a reliable estimation method is expected to yield good predictions for values in that range. In all experiments, we generate samples of 100 observations and successively omit from 10 up to 90 observations in the middle of the sample, with an increment of 10 in each step. For each of experiment, we run the programs based on Bayesian algorithms and save the estimation and prediction results. These are summarized in a series of tables and graphs in Appendix B.

Note that our goal here is not to produce diagnostic tests of misspecification, but rather to check how well the different Bayesian methods perform at estimating a correctly specified model. Hence, we assume the same form for the econometric model and the underlying data generating process (DGP) in each experiment. Since predictions are calculated as the sample means of the augmented dependent variable values generated from MCMC draws, they are generally less variable than the actual values simulated for the needs of the experiments. We can think of the Bayesian predictions as being the average of the simulated dependent variable values if the Monte Carlo experiment was repeated several times.

Since MCMC draws, such as those generated by the Gibbs sampler, are correlated by construction, the sample variance does not provide an accurate estimate of a parameter variance. Carlin and Thomas (2000, Chapter 5, pp. 170-172) suggest a correction method for variance estimation in the presence of positively correlated MCMC draws. However, this formula cannot be applied in the present framework because the Gibbs sampling draws are generally negatively correlated. Instead, we make use of another of their suggestions. Sample variance is rather calculated based on draws at specific increments (every 100 iterations) in order to reduce correlation among them.

According to Carlin and Thomas, the Normal approximation holds even in the Bayesian framework. In consequence, regardless of which of the above variance estimates is used, a confidence interval of the parameters conditional expectation might be calculated in the conventional way. Since predictions are computed through the data computation stage, as additional model parameter estimates, this holds for them as well. Student-t statistics can be derived from the same principle. Critical values for these tests are drawn from the standard normal distribution with m degrees of freedom. In our case, m represents the number of successive batches of length k used to compute the variance estimates (N-n=mk, with k=100). Since several iterations are performed for each Gibbs sampling experiment, z_{∞}=1.96 is satisfactory for a test at the 95% confidence level.

5.1.1. Simulations involving single equation models

Table B.1 of Appendix B presents the estimation results for the first series of Monte Carlo experiments based on a simple linear regression model. Although their precision generally decreases with the length of the prediction spell, all parameter estimates are significant at the 5% level, except the intercept which is not significant at conventional levels when 90 observations are removed from the sample. In most cases, real values fall within one standard deviation of the parameter estimates. In all cases, the 95% confidence interval contains the real coefficient values.

Figure B.1 shows that, compared to real values of the dependent variable, predictions are stable and accurate, even for large prediction intervals. Figure B.2 provides a typical illustration of the predictions' dispersion. It illustrates the accuracy of predictions with respect to the observable path when 50% of the sample information is missing. Except for a couple of rather extreme values, generated values of the dependent variable are contained in the 95% confidence interval around the predictions. The prediction error, measured as the difference between real and predicted values, is depicted in Figure B.3. As expected, the figure shows that the errors are smaller for shorter prediction intervals. Otherwise, the prediction errors' path display no particular pattern: errors are distributed symmetrically around the zero mean.

For the second series of Monte Carlo experiments, we started with the simplest case of a single (one- or four-period) lagged dependent variable, and successively increased the number of lags included as explanatory variables in the model specification. Table B.2 displays the results obtained by estimating a linear regression model involving a constant term and a four-period lagged dependent variable. Estimation results for the model involving a one-period lagged dependent variable are not provided because they were very similar to the preceding ones.

All parameter estimates are significant at the 5% level, except the estimate of ρ which is only significant at the 10% level for 80 omitted observations, and looses its significance (at conventional levels) when the number of missing observations is further increased. The relatively large estimate of the error variance indicates the imprecision of the estimation technique for 90% of missing information. Real parameter values all fall within the bounds of one standard deviation from the Bayesian estimates.

Despite the relatively good precision of the parameter estimates, the sampling based method does not yield very accurate predictions. The comparison with the generated data provided in Figure B.4 reveals that predictions are much less variable than the actual series. While the latter displays fluctuations of large amplitude, the former appears almost completely flat. As expected from a stationary process, since the variations in the dependent variable are marginal at the beginning of the prediction period, they tend to vanish thereafter. In fact, the prediction path varies much less than, but in phase with, the real values' trend, as illustrated in Figure B.5. Since the predictions are not very precise, their 95% confidence interval generally encompasses the real values of the dependent variable.

These observations are in accordance with the above remark regarding the way Bayesian predictions are calculated. This was also expected from the underlying dynamics, and the removal of relevant mid-sample information. Given the relative insensitivity of the predictions, the prediction errors displayed in Figure B.6 follow a similar path to the original data one. They however oscillate around the zero mean, which indicates that the predictions correctly average the simulated data.

Note that when we set the true value of the lagged dependent variable to a smaller value, such as 0.1, its parameter estimate does not show up to be significant at conventional levels. However, this is not particular to that type of model: Monte Carlo experiments have shown that, in general, coefficients associated with explanatory variables which only have marginal effects are difficult to estimate by using the sampling-based techniques considered in this study. Also, the prediction performance does not improve when a more extended lag structure is allowed for in the model specification. We obtained very similar results by successively including from one up to six lags of the dependent variable in the regression. This is why these results are not reported in the appendix. Remark that since the impact of past values on the current one is generally smaller the more distant these values are from each other, it is harder to estimate accurately coefficients at long lag length.

The third series of simulations involves a linear regression model with an explicit structure of autocorrelation of order four for its error terms. Except when 80% or more of the information is removed from the estimation sample, parameter estimates reported in Table B.3 are generally precise and accurate. The model parameters are significant at the 5% level and the true values lie within one standard deviation of the parameter estimates. The estimated error variance is generally larger than its assumed value, however. Autocorrelation parameters ϕ_1 and ϕ_4 are estimated quite accurately. Of smaller amplitude (in absolute terms), ϕ_2 and ϕ_3 estimates are not significant at conventional levels, but have the expected sign, and their true values fall within one standard deviation from the parameter estimates.

According to Figure B.7, compared to real values, predictions seem remarkably accurate. Furthermore, the 95% confidence interval generally embeds the true generated values of the dependent variable, as illustrated in Figure B.8. However, it is apparent from Figure B.9, that predictions are more dispersed and less in accordance with real values of the dependent variable at the end of the prediction interval than at the beginning. Prediction errors are also smaller on shorter intervals.

5.1.2. Simulations involving simultaneous equations models

In order to check the performance of sampling-based techniques applying to systems of equations, we consider models involving two equations. Table B.4 presents the estimation results for the series of Monte Carlo experiments relating to the SUR model. Model parameters are significant at the 5% level in all simulations, except the intercept in the first equation which is only significant at the 10% level when 90 observations are removed from the estimation sample. Except possibly when 20 observations are missing, the real value of the parameter lies within one standard deviation of the estimate for each model parameter. It is always contained within the 95% confidence interval. The covariance between the two error terms is never significantly different from zero at conventional levels, but variance error terms both are. Furthermore, the one standard deviation intervals around the variance parameters contain the true variances, except that of the second equation with 20 missing observations. The 95% confidence interval, however, includes it.

Predictions of both dependent variables, given in Figures B.10 and B.11, seem relatively accurate although less variable than the corresponding observable path. In terms of dispersion, the true value of Y_1 is more likely to be contained within the one-standard deviation from the prediction, or the 95% confidence interval, than that of Y_2, which is generally more volatile. Figures B.12 and B.13 provide typical illustrations of the precision of predictions of Y_1 and Y_2, respectively, when 40 observations are removed from the middle of the estimation sample. Predictions of Y_2 are also more frequently out of phase with respect to the observable path than those of Y_1. Prediction errors on Y_1 are smaller and symmetrically distributed around the zero mean for relatively short prediction spells, but dispersed and more important at the end for longer intervals, as depicted in Figure B.14. Based on Figure B.15, the SUR model underestimates values of Y_2 on short prediction intervals, but overestimates them on longer spells. Overall, the Bayesian estimation of SUR models seems to yield very adequate predictions.

Table B.5 presents the estimation results for a pure vector autoregressive model involving only one lag of each dependent variable in each equation (PVAR(1)). Large estimated variances of error terms indicate a convergence failure when 80 observations or more are removed from the estimation sample. Apart from these limiting cases, model parameters are generally significant at the 5% level, except possibly the first equation's intercept which is significant at the 10% level when 70 observations are omitted from the prediction sample, and some coefficients relating to the lag of the other dependent variable. Parameters on own lags are highly significant and estimated relatively accurately, while it is generally harder to estimate cross-effects.[1] Nonetheless, the cross-equations coefficients have the expected sign and are of the expected order of

[1] When we hypothesized smaller cross-effects (in absolute values), we obtained non significant results. But this does not only apply to this particular type of model: as outlined above, it is generally difficult to estimate small parameters accurately by using sampling-based techniques.

magnitude, although smaller, in absolute terms. Again for 70% of missing information or less, parameters of the error variance-covariance matrix are estimated quite accurately.

As in the univariate case of a pure autoregression, predictions of the PVAR(1) model do not vary enough to meet the fluctuations in the observable trends. This is shown in Figures B.16 and B.17. In spite of that, the true dependent variable values fluctuate around the more stable trends displayed by the corresponding predictions. Variations in predictions are more accentuated at the beginning of the prediction interval, and tend to vanish at the end. The observed Y_1 series is almost always contained within the one-standard deviation interval from predictions but sometimes exceeds it (e.g. Figure B.18), while the real Y_2 series often crosses-over the corresponding interval from below (e.g. Figure B.19). Both series are nevertheless embedded in the 95% confidence interval around predictions. The lack of precision of one dependent variable's prediction carries over to the other. Figures B.20 and B.21 corroborate the fact that Y_1 is actually being overestimated, while Y_2 is underestimated.

Complementary simulations have shown that the prediction performance of the PVAR does not improve when longer lag structures (up to order 6) are accounted for. The generated series remain more variable than the predicted ones. For lack of space, we do not provide these additional results. Note that as in the univariate case, the precision of parameter estimates decays with the lag length, as past effects on current variables become negligible.

Simulation results for a generalized vector autoregressive model of order 1 (GVAR(1)) are provided in Table B.6. Large error variances indicate convergence failure for 90 missing observations. For smaller prediction spells, the model parameters are generally significant. Parameters associated with exogenous variables are estimated quite accurately (except the second equation's intercept which is slightly underestimated), as well as those on first own lags. In contrast, as for the PVAR, it is more difficult to obtain accurate estimates of the cross effects. Indeed, the parameter corresponding to the lag of Y_2 in equation 1 is underestimated, although it is significant at the 5% level up to 60 missing observations. Similarly, the coefficient of the lagged variable Y_1 in equation 2, which is negative, is overestimated, while it is significant at the 5% level up to 40 missing observations. The error variance-covariance parameters are estimated accurately, except for the smallest sample size.

Introducing stochastic variables into the VAR model formulation clearly improves the model fit, as show Figures B.22 and B.23. Now, the predictions follow more closely the observable trends in the dependent variables and display similar fluctuations, although they are of smaller amplitude. As expected, their accuracy worsens with the stretching of length of the prediction interval. Figures B.24 and B.25 provide typical illustrations of the dispersion of predictions, together with the real dependent variables' paths. Real values almost never cross the 95% confidence interval around predictions and when they do so, it is by a very narrow margin. The distributions of prediction errors displayed in Figures B.26 and B.27 seem to indicate an underestimation of both dependent variable values. But in fact, the paths are relatively regular when we make abstraction of the limiting cases with 80% or more of missing information.

For lack of space, we do not provide Monte Carlo results for longer lag structures. Note that the general conclusions derived above also hold for more extended GVAR(p) models. Similarly to PVAR(p) models, however, the precision of parameter estimates decays with the lag length. In fact, as these usually have marginal effects on the dependent variables, predictions do not vary that much with respect to p in the GVAR(p) model setting.

Overall, the different Monte Carlo experiments performed in the framework of this empirical analysis proved that the Bayesian method is very promising for our practical application. With the exception of purely autoregressive models whose values appeared to be particularly hard to extrapolate, we obtained very satisfactory results, both in terms of estimation and prediction. Now that the sampling-based techniques have been tested in a controlled environment, let us examine how they perform in practice.

5.2. Model specification

In this section we examine, one by one, the exogenous variables whose explanatory power has been tested in each equation specification. As private vehicle use varies widely by vehicle type, cars and light trucks/vans are always processed separately. While the dependent variables are derived directly from the surveys, the independent variables are drawn from external sources of information.

Most of the exogenous explanatory variables were taken from Statistics Canada's CANSIM disk databank, while others were drawn from files kindly provided by Natural Resources Canada (NRCan) and Environment Canada (EC). Other explanatory variables were found on the Bank of Canada's and the U.S. Federal Reserve Bank's Web sites. Potential predictors were identified through a parallel macroeconomic study of the private transportation sector (see Boucher, 1998b).

Among the explanatory variables that were identified, some were not available for the entire target period, and others did not exist at the required quarterly aggregation level. Variables reported on a monthly basis had to be aggregated to the quarter level by calculating sums or averages of the values for the three months in the quarter. Others were reported only annually, so the annual value had to be repeated for each quarter of a given year in the model formulation. For lack of taking into account seasonal fluctuations in the transport aggregates, these variables give an indication of their average annual trends.

Determination of the specifications for each equation required a long preliminary selection process. The explanatory power of each model formulation was tested by calculating the ordinary least squares (OLS) estimates based stacked samples from both surveys, without consideration for the missing values in the intervening period. Particular attention was paid to selection of the best predictors, and so ignored any multicollinearity among them. Additional diagnostic tests will be conducted in the next chapter to determine if alternative specifications cause any changes in predictions.

The basic specification for each dependent variable, with only right-hand side exogenous explanatory variables, is described below. Note that these specifications are the ones that were selected following a series of diagnostic tests performed in order to determine the best model for prediction purposes. This will be followed by a brief definition of each explanatory variable appearing in the formulations. The reader is referred to the beginning of Appendix C for a more detailed description of the explanatory variables, their origin and, if applicable, the way they were modified to fit the econometric context of the study. Although cars and light trucks/vans were processed separately, we finally ended up with the same kind of specification for both vehicle types:

$$\begin{aligned} S &= \beta_0 + \beta_1 \mathrm{REGIST} + \beta_2 \mathrm{VPRICE} + \beta_3 \mathrm{FALL} + \beta_4 \mathrm{WIN} + \beta_5 \mathrm{SPR} + u_1 \\ \overline{D} &= \beta_0 + \beta_1 \mathrm{PCREGIST} + \beta_2 \mathrm{FALL} + \beta_3 \mathrm{WIN} + \beta_4 \mathrm{SPR} + u_2 \qquad (1) \\ E &= \beta_0 + \beta_1 \mathrm{FCR} + \beta_2 \mathrm{FALL} + \beta_3 \mathrm{WIN} + \beta_4 \mathrm{SPR} + u_3 \end{aligned}$$

Below we give, in alphabetical order, a list of the explanatory variables in the equations, with a brief definition of each. We will then comment on the inclusion/exclusion of certain predictors from the above specifications.

FALL, WIN, SPR: Dummy variables for the Fall, Winter and Spring quarters, respectively;

FCR: On-road fuel consumption rates of cars, annual series, expressed in litres/100 kilometres;

PCREGIST: Per capita registrations, based on the ratio of the number of passenger automobiles registered in Canadian provinces to the number of people of driving age (16 years or more);

REGIST: Annual number of passenger automobiles registered in Canadian provinces. The passenger automobile category includes light trucks and vans used for private purposes. It is thus defined in opposition to commercial vehicles and can be used in modelling variables for both light trucks/vans and cars;

VPRICE: Average price of new vehicles, on a quarterly basis, in constant dollars of 1986;

Note that the selected formulations for each equation involve the annual figures currently used by NRCan. We also considered an alternative specification from the transportation literature, which involves income and fuel prices in all equations (e.g. Dahl and Sterner, 1991). Alternatively, annual real personal disposable income (ARVDPC) and quarterly national income (QNATINC), both in constant dollars of 1986, were used as measures of revenue. As indicators of fuel prices, we successively considered the quarterly series of a composite consumer price index of different fuel types (QFCPI), and an average of unleaded fuel prices (QUNLFP), both in constant dollars of 1986. As will be shown in the next chapter, the original specifications yield better results, as far as predictions are concerned.

During a preparatory stage to the modelling process, it was noted that quarterly fuel sales (QFSALES) would be a better predictor for the total distance travelled than annual registrations. On the one hand, quarterly retail sales of gasoline had been shown to be an effective support for the series on total distance travelled by automobiles. However, it was finally determined that annual passenger automobile registrations captured the overall trend of the series more effectively. Combined with other explanatory variables that factor in seasonal variations in the series requiring prediction, registrations were thus preferred to fuel sales. Contrary to what one might expect, the correlation between quarterly gasoline sales and per capita registrations was not too strong, allowing for both variables to enter the average distance formulation jointly.

Similarly, annual fuel sales to light trucks and vans had initially been identified as a good predictor for the total distance travelled equation by this type of vehicles. Indeed, NRCan had provided us with the series for annual fuel consumption by vehicle type.[2] Since gasoline is not the main type of fuel used by light trucks and vans, quarterly gasoline sales would not be as good as an explanatory variable in this equation. For light trucks and vans, but not cars, the annual fuel consumption series yielded better results than quarterly gasoline sales. However, NRCan did not accept our request for an update of the series ending in 1995, so we were unable to test it as a predictor for the distance-travelled-by-light-trucks/vans equation. Registrations were thus selected for this equation too.

During the preparatory study, it was agreed, for the sake of convenience, that we would produce a model for total distance travelled by all vehicles, instead of the average distance travelled per vehicle. At first glance, it seemed easier to explain changes in the former than in the latter. Our plan was to deduce posterior predictions for average distance by dividing predictions of the total distance by those of the average vehicle stock. However, further tests showed that, in the resulting predictions of average number of kilometres driven, seasonal variations were much too low.

This may be explained by the fact that the models developed for the average number of vehicles and total distance equations were very similar. In particular, variables capturing quarterly variations, such as seasonal dummy variables, would generate effects that would tend to cancel out one another. Furthermore, preliminary standard tests based on stacked data from both surveys indicate that the total distance series

[2] The data compiled by NRCan present an interesting alternative in this particular case. In addition to disaggregating consumption by vehicle type, NRCan calculates total light truck/van consumption of all kinds of fuel used by this vehicle type. Data on fuel consumption by vehicle type come from the Trans.xls Excel file, which NRCan uses for preparing its annual publication.

might be integrated, while the average distance series are not. Consequently, the decision was taken to model average distance directly.

In order to capture quarterly fluctuations in each series requiring predictions, seasonal dummy variables and quarterly averages of Canadian temperatures (in degree-days) were alternatively considered. The Summer quarter, being the reference category for the first set of dummy variables, was excluded from the model specification. In all cases, the seasonal dummy variables were shown to capture quarterly variations more effectively. Note however that the temperature variable could not be tested as a predictor in the specification of equations regarding cars, as the series starts in 1981.

Interest rates (IRATES) were tested in the vehicle stock equations because, as pointed out in Section 2.4 of Chapter 2, a privately owned vehicle is a durable good whose purchase should be considered as an investment. We considered interest rates on personal consumer loans as a proxi for those on private vehicle purchases. However, this series could only be used as a regressor in specifications regarding light trucks and vans, because it starts in the fourth quarter of 1980. For cars, we tried a similar series of U.S. interest rates[3] and a quarterly average of Canadian interest rates on prime business loans from chartered banks.

Another exogenous variable whose explanatory power has been tested in the ownership and use equations is the quarterly gross domestic product (GDP) of the public transit sector. This aimed at accounting for the substitution between private versus public transport modes. Fixed costs associated with owning a private vehicle, as opposed to variable costs reluctant to using such a vehicle, were tested in the ownership and use

[3] The U.S. and Canadian interest rates on consumer loans move together, given the interdependency of the two economies. Testing based on the number-of-light-trucks/vans equation showed that the effectiveness of the two series as explanatory variables was about the same.

equations, respectively. These variables were computed as quarterly averages of relevant CPIs.

5.3. Estimation results

Estimation and prediction results obtained from the application of the Bayesian method to the econometric model (1) are displayed in the supporting Appendix C. Although parameter estimates are only of secondary interest, comparing estimation results from different model types and specifications sheds light on the contemporaneous and dynamic relationships existing among variables under study. It also permits the evaluation of the relative performance of competing MCMC estimation techniques based on distinct prior assumptions. This provides some leads to determine the best prediction model.

Reporting all estimation results produced in the framework of this analysis would be too space consuming. We only tabulate results for the selected model types and specifications,[4] and comment on divergences obtained with alternative formulations. Arguments to support our model choice will be provided in the next chapter. Specifically, the generalized vector autoregressive model (GVAR(1)) involving, in addition to the stochastic and deterministic variables listed in the preceding section, a one-period lag of each dependent variable, is shown to yield the most reliable and accurate predictions. Parameter estimates for such a model are presented in Tables C.1-C.3 of Appendix C, together with their standard errors and Student-t statistics. CPU times required for the estimation of each model are also reported in Table C.3.

Given the large number of potential different specifications, only statistically significant exogenous variables are kept in the basic model formulations. Consequently,

[4] Estimation results for other model types and specifications may be obtained from the author upon request.

almost all variables figuring in the simple linear regression models have strong explanatory power. Seasonal dummy variables are the exception in equations regarding the average numbers of vehicles. While these variables are not significant (at conventional levels) in the cars equation, solely the Fall dummy variable is significant (and only at the 10% level) in the specification for light trucks and vans. Its coefficient is negative, indicating that the light trucks and vans stock is larger in the Summer than during the Fall season.

This is consistent with both people's tendency to garage part of their vehicles fleet in arid seasons, and to buy vehicles for private use in seasons with milder climates. Given that cars have less robust characteristics, we would expect these effects to have a larger influence on their stock, however. Figures might be mitigated by the inclusion of minivans, whose features and use resembles more closely that of cars, in the light trucks category. Since the vehicle stock is likely to vary from one quarter to the next, and given that no other explanatory variable satisfactorily accounts for such fluctuations, we felt that seasonal dummy variables should be kept in the model specification despite their lack, or poor degree, of significance.

First attempts to include some dynamics into the basic simple linear regression framework provide the following conclusions. Introducing a lagged dependent variable as a regressor into the model specification mainly affects the estimates of the average number of light trucks and vans equation. In that case, the coefficient of the one-period lagged dependent variable is nearly 0.8 and highly significant.[5] Once this variable is added to the model specification, the parameter on passenger automobile registrations is no longer significant at conventional levels. The resulting predictions follow a similar path as their simple linear regression model counterpart, but seasonal fluctuations are less

[5] In alternative model specifications involving income and fuel prices, however, the corresponding parameter turned out to be small (about 0.3) and not significant.

accentuated. For other equations, the additional parameter is likewise positive, relatively smaller and not significant. Hence, predictions are not significantly altered.

Within more extended GLAG(p) structures, parameters associated with longer lag lengths are generally unsignificant and smaller than those on first lags. In spite of that, predictions of the average number of light trucks and vans vary considerably with the allowed maximum lag length, increasing at a faster rate for larger values of p. Statistical criteria, such as those proposed in the next chapter to determine the appropriate order of an autoregressive process, suggest that the inclusion of a single lagged dependent variable should be sufficient to model the light trucks and vans fleet size accurately, provided that other exogenous factors are also accounted for in the estimation process.

Predictions of the PLAG(4) model, involving a single four-period lagged dependent variable, also differ considerably from the previous ones. Although all LAG models' predictions follow the same trend, seasonal fluctuations resulting from the PLAG(4) estimation are more accentuated, in conformity with those of the linear regression model (LIN).[6] In spite of that resemblance, the coefficient of the fourth-order lagged dependent variable is relatively large (α_4=0.517) and highly significant.

Which of the PLAG(4) or the LAG(1)[7] models better fits the light trucks/vans stock data is unclear, although the argument developed in Section 4.2.2, while introducing the econometric treatment for LAG models, favours the use of the LAG(1), which accounts for the fact that the vehicle stock in a given quarter is determined as the stock in the preceding quarter plus a residual term. Recall that the remainder equals the difference between new vehicle sales and old vehicle sent to scrap. Furthermore, since we are conditioning on the first p observations, using lags of shorter lengths results in an

[6] The PLAG(4) yields even more volatile predictions than the LIN at both ends of the prediction interval.

[7] For an obvious reason, we remove the capital "G" letter in front of the GLAG(p) notation when we refer to the model involving a single lag of order 1, and write LAG(1) instead.

improvement in the number of degrees of freedom. Hence, at least from a theoretical point of view, the LAG(1) model should provide more reliable results.

An alternative, though somewhat restrictive,[8] route to capture the underlying dynamics in single equations is to introduce an explicit autocorrelation structure of order four (AR(4)) for the model's error terms. At least for order one, autocorrelation parameter estimates are significant in both vehicle stock equations. Note, however, that the estimate of ϕ_1 for light trucks and vans is almost twice as large as that for cars. Accordingly, predictions of the AR(4) model disagree with those of other models more markedly in the former case. The average distance and weighted fuel consumption equations seem less adversely affected by serial correlation of either type.[9]

In order to account for possible interdependencies among the three transport aggregates, simultaneous equations models have also been considered. Since it represents the simplest form of multidimensional model, relying only on contemporaneous and exogenous information, the SUR model has first been estimated. Compared to its single-equation counterpart, the simple linear regression, SUR estimation generally improves the precision of all parameter estimates. However, coefficient estimates themselves do not change that much (apart possibly for the intercept) neither in sign nor in magnitude. Provided that some estimated off-diagonal elements of the error variance-covariance matrix are significant in the cars model, it is justified to estimate the system of equations as a whole. Predictions of the average

[8] Remember that the explanatory variables matrix does not involve any lagged dependent variables here. As the analysis of the estimation results will show, once an extended lagged structure is accounted for explicitly in the model formulation, there is no remaining autocorrelation in the residuals.

[9] In the alternative model specification for the light trucks and vans fuel efficiency, the four-period lagged dependent variable coefficient was found to be large (>0.6) and statistically significant. Likewise, the first-order autocorrelation parameter estimate is larger than 0.5 and significant.

number of vehicles (of both types) are higher for the SUR model than for single-equation models, especially at the beginning of the prediction interval.[10]

As already highlighted in Chapter 4, predictions could also be derived from the estimation of an unrestricted reduced form (URF) of the system of equations. The estimation method developed in Section 4.3.1 is applicable to the URF because it can be expressed in the form of a SUR model. We have estimated an URF of our structural model using the Gibbs sampling algorithm with data augmentation. Since registrations and per capita registrations are highly correlated, they could not be both incorporated in the URF. We chose to leave out the second variable because the former one better explains the number of vehicles, which is the hardest to predict among the three dependent variables.

In terms of predictions, we obtained very close results from the SUR and URF estimation. Given that it is easier and more compelling to interpret the estimates derived from the structural form of the model, and also because these can be put into correspondence with more restricted estimates obtained in the univariate case, we decided to keep focussing on the structural form of the system. Nevertheless, we also report on the URF predictions for comparison purposes in the next section.

When an extensive autoregressive structure is allowed in the form of a pure vector autoregressive (PVAR) representation, results change drastically. Except for the deterministic variables (and even the constant and seasonal dummy variables are not all significant in the vehicle stock equations), only coefficients on the first own lagged dependent variables are significant. Their estimates are positive in all cases. In spite of

[10] Predictions of the SUR model differ markedly according to the model specification (registrations-vehicle prices versus income-IPC settings). Not only the gap between those predictions is large, but also seasonal fluctuations appear to happen at different episodes. The selected model specification brings predictions better in accordance with those of alternative model types.

that, results vary according to the lag length. We include up to six lags of each dependent variable in each equation. Predictions of the vehicle stock increase with the number of lags included in the model, with those at highest lag lengths falling more in the range of other models' predictions. As expected, vehicle ownership is positively correlated with the average distance travelled and negatively correlated with the fuel consumption rate. The latter two variables appear to be negatively correlated as well. Off-diagonal elements of the variance-covariance matrix are not significant, however.

Reintroducing stochastic variables in the VAR setting clearly improves the model fit. Indeed, most of the variables having an explanatory power in more restrictive formulations remain significant in the GVAR model. Furthermore, coefficients on first own lagged dependent variables continue to be significant. For comparison purposes, we wished to include the same number of lags in the GVAR and PVAR models. We consider up to six lags of each dependent variable in each equation relating to cars, but only five in those regarding light trucks and vans, due to a smaller sample size. Providing that other parameters on lagged variables are relatively small and not significant, the exact number of lags included in the GVAR model has little impact on the resulting predictions. Nevertheless, adding more than one-period lags impairs on the estimations' precision, as will be argued in the next chapter. Once having explicitly accounted for the relationships among the three dependent variables, all the off-diagonal elements of the error variance-covariance matrix become unsignificant.

Detailed comparisons of predictions obtained from the diverse model types are provided in the following section.

5.4. Prediction results

In this section, we comment on the predictions resulting from the application of the Bayesian methodology outlined in Chapter 4 to the estimation of the simultaneous equations models described in the preceding sections. By default, the reported predictions come out of the Bayesian estimation of the selected model, i.e. GVAR(1). Tables C.4 and C.5 of Appendix C provide the completed (with predicted values) time series for variables relating to cars and light trucks/vans, respectively. The shaded area corresponds to the prediction period, while the remaining entries relate to estimates derived from survey data. Recall that survey-based estimates were previously submitted to the series of adjustments, described in Chapter 3, aiming at improving their compatibility. Predictions are reported together with a 95% confidence interval to evaluate their relative precision.

Graphical comparisons of predictions obtained from the various models types considered in Chapter 4 are also provided, and the supportive argumentation of methodology developed. Apart for the PVAR model, which only involves lagged dependent variables, and the URF formulation, which does not account for per capita registrations, all model specifications include the same exogenous variables as the GVAR, i.e. those described in Section 5.2. That way, prediction results are easier to compare and contrast. See Appendix C for figures and tables pertaining to the prediction results. Note that all figures appearing in Appendix C are conceived in such a way as to reflect the discrepancies in predictions of different model types, if any. The number of lags of the dependent variables appearing in autoregressive models is selected in accordance with the argumentation developed in the preceding section.

Two different methods of estimation were used for each type of model. While only the results of the Bayesian application are reported, a degenerated method[11] was also

[11] The degenerated method involves estimating, for each stage of the iterative process constituting the sampling method, the parameters of the model by using standard econometric methods instead of the Bayesian estimators. Its name refers to the fact that, assuming natural conjugate priors in the simple linear regession framework, standard estimators result from imposing degenerated values to the prior distribution parameters. While it has no specific theoretical basis, it is known to produce "good" results in practice.

considered for validation purposes. Whenever the two estimation methods yielded similar predictions,[12] the results were considered valid. When they did not, further investigation was required. Section 2.5.2 has provided regularity conditions under which different MCMC algorithms converge. In practice, these might be difficult to verify. We will return to this issue in the next chapter. For the moment, all estimations and predictions were derived from the application of the appropriate sampling-based algorithm, iterated N=5000 times, and computed using results from the last 4000 iterations following the burn-in period (n=1000) to reduce the effect of the choice of the starting values.

5.4.1. Equations for the average number of vehicles

Figures C.1 and C.2 illustrate, respectively, predictions of the average number of cars and light trucks/vans derived from the estimation of the various types of models considered in Chapter 4. Since the series to be completed show a relatively stable growth rate, predictions increase monotonically throughout the prediction period. The amplitude of seasonal fluctuations varies more with respect to vehicle type, especially during the NaPVUS period. The number of light trucks and vans grew at a faster rate than the car stock during the target period. This observation is based in large part on the inclusion of minivans in the light trucks category.

Figure C.1 shows that predictions of the number of cars change very little from one model to the next, despite their different specifications and prior assumptions. This observation upgrades our confidence in the robustness of the results. The main discrepancies arise from the URF, whose almost flat predictions do not seem to mimic

[12] In some cases, although estimates of the model parameters differ considerably depending on the estimation method, predictions calculated with these methods are very similar. We must therefore conclude that there is an adjustment feature in the overall estimates that causes the methods to produce convergent results, in terms of predictions.

correctly the seasonal fluctuations observed in the remaining of the survey-based data. Accounting for serial correlation within the series under study yields more regular seasonal variations through the prediction spell. Allowing for the inclusion of stochastic independent variables within the VAR model specification, in addition to the lagged dependent variable structure, provides a way of capturing macroeconomic fluctuations as well. Finally, estimating the system as a whole, instead of equation by equation, slightly increases the predicted number of cars for the intermediary period, between the two surveys. This is consistent with the expected positive effect of distance travelled on car ownership. Therefore, predictions for the cars stock equation seem reliable, although its specification will be further questioned in the next chapter.

The analysis is very different for the light trucks and vans series. As illustrated in figure C.2, the estimated number of light trucks and vans varies considerably from model to model. Although to a lesser degree, this finding can easily be generalized to all the variables requiring predictions: since the prediction period is longer, and the estimation basis shorter, for light truck/van variables than for car variables, the latter are likely to be predicted with more accuracy. In addition, because estimates of car variables are based on larger samples, they form relatively more stable aggregate series than the light truck/van estimates. As a possible consequence, the cars stock estimates also fluctuate more regularly from quarter to quarter than their counterpart.

The average number of light trucks and vans has increased dramatically during the prediction period. Seasonal variations in the estimated number of light trucks and vans are also more accentuated during the NaPVUS' sampling period than in the FCS', thus compounding the difficulty of predicting the series during the intervening period. These observations can be explained, in part, by the recent increased popularity of minivans and sports utility vehicles which are used more intensively in arid seasons with rigorous climates.

Firstly, considering the single equation estimation results shows that introducing lagged dependent variables as regressors, or assuming that the error terms follow an AR(4) process, considerably modifies the path of the predicted values. Note that the alternative ways of incorporating some dynamics in the linear regression framework have opposite effects on predictions, especially at the end of the prediction period. Surely, accounting for serial correlation is crucial, but also tricky, in this case.

Secondly, a comparison of predictions with the remainder of the series reveals that a model solely based on past experience, such as the PVAR model, is likely to understate the average number of light trucks and vans, especially at the end of the prediction period. Indeed, the huge gap between the last predicted value and the first observation from the NaPVUS is very unlikely. This observation holds no matter how many lags of the dependent variables are included in the model formulation. Indeed, predictions of the PVAR(p) model increase when p gets larger. The tendency of PVAR models to underestimate dependent variables at the end of prediction intervals has already been outlined in the interpretation of Monte Carlo experiment results. Note that the predictions resulting from the linear regression model with AR(4) errors display a likewise, but even more important, gap at the moment they join end the NaPVUS data.

Predictions of the other types of simultaneous equations models (the URF, SUR and GVAR(1)) agree much more easily. The GVAR(1) predictions are slightly larger than those of the SUR and URF. This is consistent with the estimated positive and significant effect from the stock in the previous quarter, and with the acknowledged fact that North American countries approach car saturation. The increasing rate of stock accumulation observed at the beginning of the prediction period in three completed series can be justified by a larger representativeness of the light trucks and vans in the overall vehicle fleet starting in the late 1980's. As will be shown in Chapter 6, the SUR and the GVAR(1) models are the main competitors in the search for the best prediction model type.

The importance of obtaining reliable predictions of the private-use vehicle stock series cannot be overstated. Because it is the only total quantity variable in the simultaneous equations model, the predictions of other variables indicating overall use of private vehicles, such as the total distance and fuel consumption, are dependent on it. Developing a reliable profile of private vehicle use is contingent on accurate predictions of the number of vehicles. For this reason, the next chapter will focus on the testing of the model formulation for the average number of vehicles.

5.4.2. Equations for the average distance travelled per vehicle

Figures C.3 and C.4 compare predictions of the average distance, based on the estimation of different model types for cars and light trucks/vans, respectively. Once again, the URF predictions distinguish themselves by displaying less intense seasonal fluctuations. Otherwise, the predictions evolve in the same way as the remainder of the series, following the same slightly upward trend and comparable seasonal variations. Since the average distance predictions change little from model to model, they appear to be relatively robust.

Figure C.5 contrasts predictions of average distance travelled by vehicle type. On average, light trucks and vans cover longer distances than cars. Both upwards and upwards variations in average kilometres driven by quarter appear more pronounced for light trucks and vans than for cars.

5.4.3. Weighted fuel consumption rate equations

As shown in figures C.6 and C.7, the weighted fuel efficiency predictions for cars and light trucks/vans, respectively, vary little among the different model types we estimated. Now, the main difference is observed from the PVAR(p) predictions for light trucks and vans, which are higher than those derived from other models. Although only the results of the PVAR(6) estimation are drawn, predictions from more restricted lag structures were even higher.

In spite of that, predictions of all models follow the same decreasing trend as the remaining estimated series. In general, seasonal variations in the survey data are also passed along to the prediction period. Since predictions change very few according to the hypothesized model type and underlying priors used to obtain them, we are relatively confident in their robustness.

Weighted fuel consumption rates of cars and light trucks/vans are compared in Figures C.8. As expected, cars are more efficient than light trucks and vans. Furthermore, the gap between the two series is relatively stable, although it seems slightly wider at the beginning of the period. The data reflect manufacturers' serious efforts to bring about a continuous improvement in light vehicles' fuel efficiency and the progressive disappearance of large cars from our roads. Both curves follow a similar path, with variations of about the same amplitude, indicating that seasonal variations in fuel efficiency are satisfactorily factored into the prediction models.

Chapter 6

Reliability and Robustness of the Results

Chapter 5 and the supporting Appendix C have shown that predictions do not vary that much with respect to the postulated model and the applied sampling-based estimation techniques. The equation for the average number of light trucks and vans is one exception. The irregularity of the series and its accelerating trend and seasonal fluctuations in the mid 1990's impair on its completion. We have already provided some explanations, but would like to pursue this further.

The present chapter looks at estimation results in more depth, and from a different perspective. We first propose a more effective way to discriminate among different model formulations according to their ability to predict actual numbers on small intervals surrounding the prediction spell. A comparison of the estimation results is established in order to determine the best prediction model. That choice being made, we discuss the reliability and robustness of the results with respect to the underlying estimation method and prior assumptions.

Section 6.1 proposes a statistical criterion to compare the prediction power of competing model specifications. According to this criterion, the best model form and model specification are determined in Sections 6.2 and 6.4, respectively. Special attention is devoted to the light trucks and vans stock equation. Since the model ultimately selected involves lagged dependent variables, other statistical criteria are used, in Section 6.3, to determine the exact number of such variables that should be included in its formulation. Robustness checks are then applied on the selected model. Convergence of the MCMC iterative process is assessed in Section 6.5. Section 6.6 checks the sensitivity of the Bayesian estimates with respect to the prior choice.

6.1. Evaluating prediction performance

Different model formulations and applications of MCMC estimation techniques can lead to discrepancies in the resulting predictions of the variables of interest. Notably, the estimates of the average number of light trucks and vans are highly sensitive to the postulated econometric model. We need a way to determine which prediction model is preferred. We focus on the light trucks and vans stock equation to perform this exercise because it provides the most discordant results.

We could develop a series of diagnostic tests to compare the different model specifications. In our Bayesian framework, these would rely on posterior odds ratio (POR). POR tests (Jeffreys, 1961) involve taking a decision based on the ratio of posterior probabilities associated with each alternative. The latter can be further decomposed as the product of prior probabilities assigned to each alternative times the expected value of the likelihood function, given that the parameter vector takes on values that are consistent with the corresponding hypothesis. The ratio of the expected likelihoods is defined as the Bayes factor (BF), a statistics which depends on the priors only through their effect in determining the posterior distributions.

Whenever underlying integrals have no close form solutions, the POR has to be calculated using numerical integration. Both Gelfand et al. (1990) and Carlin and Polson (1991) suggest a testing procedure for models whose parameters are computed via the Gibbs sampler. However, the method can be very time consuming as it involves replicating the Gibbs sampling process several times to evaluate the expected value of the likelihood function associated with each hypothesis under test. Furthermore, it requires that the normalizing constants (which make each posterior density proper, i.e. integrate to unity) be evaluated for the POR ratio to make sense. Remember that knowledge of the normalizing constant is unnecessary for estimation purpose. Hence, additional series of Monte Carlo experiments would be required to determine both proper densities entering each POR.

Mainly for reasons of computational convenience, we adopt a different strategy. Recall that the primary interest is to obtain good predictions for the transport aggregates, and not necessarily to estimate a structural model based on such variables. It thus seems natural to focus on the predictions rather than on parameter estimates. Provided that values of the dependent variables are observed within the limits of the two surveys sampling periods only, there are no benchmarks to which predictions can be directly confronted. A natural way to circumvent this problem is to check how each prediction model performs in areas, surrounding the prediction spell, for which survey information is available. In other words, we ask how each model would succeed at predicting transport aggregates if one of the surveys' sampling periods was shorter, or the interval between them was longer. This is a useful way to determine how each model fits the data within the interval delimited by the two surveys.

A statistical criterion to compare prediction performances of competing models on such enlarged intervals still needs to be selected. Just as for the specification tests described above, the comparison could be based on POR tests. Indeed, the expected likelihood of the observed values of the dependent variable can be computed and compared across different model settings, assigning equal priors to each alternative. Once again, the normalizing constants would need to be evaluated first. We adopted a more standard and less time consuming approach. It consists in comparing the root mean squared errors (RMSEs)[1] of predictions with respect to real values of transport aggregates. Since the reference (the observed value) is the same for each alternative specification, any number of models can be compared that way. Among others, Litterman (1986) and Kadiyala and Karlsson (1993, 1997) use the RMSE criterion to compare prediction performances of alternative models.

[1] The RMSE is an average measure of the discrepancies between predictions and real values. It is calculated as the root of the sum of the squared discrepancies calculated for each omitted survey observation, divided by the number of such observations.

The two following sections consider the RMSE criterion in order to determine the best model form and model specification for prediction purposes. In both cases, we consider three different scenarios. These consist in removing observations either at the end of the FCS sampling period, at the beginning of the NaPVUS sampling period, or both. Those observations are treated as the remainder of missing data and predicted in the same way. The RMSEs are calculated based on an entirely new set of Bayesian estimates since the MCMC algorithms have to be reran each time to achieve data completion. Successively from one up to eight observations are omitted from the sample basis in each scenario. Note that the NaPVUS only provides eight quarterly data for each dependent variable, so the last case is equivalent to using only the FCS information for prediction purposes.

6.2. Model selection

Tables D.1-D.3 of Appendix D show the RMSEs resulting from the removal of observations from the sampling basis of the FCS, the NaPVUS and both surveys, respectively. First note that the NaPVUS data are much harder to predict than the FCS observations. This was to be expected since the former are higher and more volatile than the latter. Generally speaking, system estimation outperforms single equation estimation, except for the linear regression model involving autocorrelated error terms which yields good results when sample information is removed simultaneously from both surveys.

It is hard to say more by looking at these tables in isolation. Ordering models in ascending order of RMSEs provides clearer conclusions. Table D.4 highlights the fact that the generalized vector autoregressive (GVAR) model is the best predictor of FCS data. Whenever observations are removed at the beginning of the NaPVUS sample, however, the SUR model outperforms the GVAR model, as illustrated in Table D.5. The results are even more mixed when observations are omitted at both ends of the prediction basis, as shown in Table D.6. In that case, the linear regression model with AR(4) errors becomes the best prediction model. The GVAR or the SUR model only comes on the second rank, depending on the number of additional observations for which predictions are required.

For each type of model, we have also graphed the predictions obtained once sample information is partially neglected, and contrasted them with the original predictions exploiting all available empirical information. Appendix D provides selected figures to support the following argument. We found that increasing the bounds of the prediction period mainly affects predictions resulting from the linear regression model including an explicit autocorrelation structure for the error terms, as expected from the underlying dynamics. Figure D.1 shows the extent of variability in predictions resulting from that model estimation when more and more observations are dropped from the NaPVUS. Predictions fluctuate even more whenever part of the FCS sample information is omitted, as shown in Figure D.2. Even the omission of a single observation results in large discrepancies in predictions obtained via this model, depending on which survey sample it is withdrawn from, as illustrated in Figure D.3.

Comparatively, predictions obtained by estimating other model types are much more stable, and diverge from the original predictions only when the number of omitted observations becomes relatively large. The precision of predictions also decreases with the number of lags of the dependent variables included in VAR types of models when the number of missing data is pushed to extreme values. This is why Figure D.4 relates to predictions of the GVAR(1) model, involving only one-period lagged dependent variables. This figure sketches typical results from the estimation of GVAR(p) models of any order p. Predictions diverge substantially only at extremely large numbers of omitted observations from the estimation basis. Otherwise, they are relatively stable. The same observations hold for the SUR model. In this sense, the two simultaneous equations models yield more reliable and robust results than single equation models.

The PVAR model which yields good results in some cases is in fact unstable, as expected from the results of the Monte Carlo experiments performed in Section 5.1 of Chapter 5. It leads to large fluctuations and general under-estimation, especially for the NaPVUS data. We conclude that it is important to use the stochastic explanatory variables, in addition to the autoregressive structure, in the modelling process. Figure D.5 provides a typical comparison of the three competing models' predictions. Note that predictions of the GVAR(1) model follow more closely those of the SUR model, while those of the linear regression model with AR(4) error terms are generally lower and more disperse.

Both the SUR and GVAR formulations provide stable results, except when the number of omitted observations and/or dependent variables lags is increased to rather large values. Furthermore, predictions resulting from these models follow relatively similar trends. Recall that estimation results displayed in Section 5.3 of Chapter 5 confirm the importance of including at least one-period lagged dependent variables into the model structure. Therefore, we continue to prefer the GVAR model to the SUR model for prediction purposes. It remains to choose the exact number of lags of the dependent variables to be included in the model specification, however. The next section proposes a simple statistical method to determine that number.

6.3. Selecting the autoregressive order in models involving lagged dependent variables

As argued in Section 5.4 of Chapter 5, predictions are relatively invariant to the number of lags, p, of the dependent variables included in the GVAR(p) model specification. Except for first own lags, estimated coefficients of these variables are generally small and unsignificant. Still, in some cases, RMSEs' ordering changes with the inclusion of more lagged dependent variables in the GVAR model specification, as shown in the preceding section. The choice of the appropriate order, p, of a GVAR(p) formulation thus remains a crucial issue for prediction accuracy.

Bayesian estimation of VAR models out predicts longer lag lengths better than unrestricted (OLS) estimation methods. Provided that we are dealing with quarterly data, considering at least four lags is advisable. In PVAR settings using seasonal data (see, for instance, Litterman, 1986; Kadiyala and Karlsson, 1993, 1997), up to six lags are considered. In the GVAR modelling framework, the reduced sample size limits this number to five for light trucks and vans.

Two reasons motivate the issue of choosing p carefully. First, Monte Carlo experiments performed in Section 5.1 of Chapter 5 have shown how difficult it was to estimate small effects accurately, and how this impairs on the predictions' precision. Second, models involving long lag length structures are more likely to suffer from multicollinearity among their regressors. Collinearity, or multicollinearity, arises when there are quasi-perfect linear relationships among some regressors. In this case, it becomes difficult to estimate individual model parameters accurately and inference is mitigated by their large variance estimates. Leamer (1978) however outlines that despite problems inherent to the interpretation of results about single parameters, linear combinations of coefficients may be estimated quite accurately even in the presence of severe multicollinearity. Since the main interest of the present study lies in the prediction of the dependent variables (which is actually a linear combination of the model parameters) rather than on the estimation of individual parameters of the system, multicollinearity is not as serious a concern as the first point raised above.

Here also, posterior odds ratio tests (POR), or the associated Bayes factor (BF) statistical tools, could help in determining the appropriate model formulation. We adopt a shortcut suggested by Schwartz (1978) to alleviate the computational burden. Schartz proposed the Bayesian Information Criterion (BIC) to compare two nested models in order to determine the right dimension of a model specification. He showed that the change in the BIC provides a rough approximation to –2 log(BF) which improves as the sample size, N, gets large. The BIC is defined as:

$$\Delta \text{ BIC}=\text{LR} - (\rho_2-\rho_1) \log(\text{N})$$

where Δ stands for the change in the statistical measure, LR is the standard likelihood ratio statistics for the test, and ρ_i represents the number of parameters in model i, for $i=1,2$, and $\rho_2>\rho_1$.

The Δ BIC thus provides some evidence of the odds in favour of the unrestricted model 2 over the restricted model 1 based solely on empirical evidence. Note that this measure is independent of the prior choice, and that it penalizes models with numerous parameters by a factor proportional to the logarithm of the sample size. Another commonly used criterion in standard econometrics is the well known Akaike Information Criterion (AIC) defined as:

$$\Delta \text{ AIC} = \text{LR} - 2\ (\rho_2-\rho_1).$$

This statistical measure also favours parsimonious specifications, but to a lesser extent than the BIC, given that it does not explicitly account for the sample size in its penalty term.

We use these two criteria in order to determine the appropriate number of lagged dependent variables to include in the GVAR specifications. Specifically, we compare models involving p and p+1 lags, varying the value of p from 1 to 6 for cars, and 1 to 5 for light trucks and vans. Since we are conditioning on the first p observations, we use the same number of observations in the estimation of each of the two models to be compared, i.e. N=T-(p+1). That way, the comparison is fairer and does not favours models estimated on the basis of larger samples right from the start.

Results from such comparisons are displayed in Tables D.7 and D.8 for models pertaining to cars and light trucks/vans variables, respectively. A negative change in one of the criterion value indicates that it is not beneficial to add up one more lag to the model formulation. In other words, the loss in the degrees of freedom in doing so exceeds the gain in the likelihood improvement. Providing that stochastic and deterministic exogenous explanatory variables are also included in the both model specifications, the inclusion of a single lag of each dependent variable is globally sufficient. This is consistent with the fact that, among all coefficients associated with lagged dependent variables, only those on first own lags are significant in the GVAR(p) estimates, even for p>1.

The advantage of using parsimonious specifications is even more important than it appears at first glance. Since we condition on the first p observations of the sample in estimating a model with p lags, considering longer lag length structures further reduces the number of degrees of freedom. This feature is particularly important in our practical application which is based on relatively small samples. Now that the general modelling framework has been established, we turn to the selection of the exogenous variables that should be included in the GVAR(1) specifications for each vehicle type.

6.4. Selecting model specification

Primary selection of the explanatory variables set was based on the predictive power of the model within the limits of available sample information. That is, OLS regressions were performed on the stacked observations from both surveys and fitted to exogenous time series reduced in the same manner. Independent variables that were found to have explanatory power were kept for further investigations. The selected model specifications for each equation have already been described in Section 5.2 of Chapter 5.

In most cases, the model specification does not affect the prediction results since the PVAR formulation yields comparable predictions. Once again, the average number of light trucks and vans constitutes the exception. For that equation, we considered several alternative specifications in order to determine the best prediction model. Section 5.2 provided a list of the independent variables whose prediction power has been tested in the stock equation.

A first selection of the most promising candidates was based on the statistical significance of each variable, this time based on Bayesian estimates. This left us with three model specifications which give relatively different predictions: a basic specification (variation 1), preliminarily selected by performing classical statistical tests based on the reduced-stacked samples, a smaller version (variation 2) of the former involving only statistically significant variables (at conventional levels) from a Bayesian viewpoint, and a completely different model specification (variation 3) involving income and prices in each equation (as in Dahl (1986) and Dahl and Sterner (1991)).

The exogenous explanatory variables included in each model formulation are listed in Table D.9 of Appendix D. Note that the table also comprises a fourth variation involving quarterly figures of national income and unleaded fuel prices. Although the estimation of variation 4 in the framework of a simple linear regression model appeared to yield good results, it leaded to too high predictions, especially at the beginning of the prediction spell. This is shown in Figure D.6 which compares the predictions obtained for each model formulation.

Moreover, the fact that we were unable to estimate a SUR model involving these two variables in each equation indicates that they are highly correlated. This explains why Figure D.6 contrasts predictions based on the simple linear regression model, instead of a less restricted formulation. We faced the same multicollinearity problem with all specifications including one of the two measures of income and fuel prices. Accordingly, we did not proceed to further tests using such formulations. We finally adopted a fallback solution which consists in using the quarterly consumer price index of composite goods (QIPC) as a proxi for fuel prices. This variables appears in variation 3.

Again, we appeal to the RMSE criterion to discriminate among the remaining three formulations. Tables D.10-D.12 show the consequences of omitting observations from the sampling basis of the FCS, the NaPVUS and both surveys, respectively. Only RMSEs pertaining to the estimation of the GVAR(1) model are reported in the tables because this model has already been shown to yield the most accurate predictions. However, we made that kind of comparisons based on all model types and obtained similar results conclusions. In all cases, variation 2 outperforms the alternative specifications in terms of the precision of resulting predictions.

In particular, private registrations and vehicle prices form a better set of predictors of the private-use vehicle stock than income and fuel prices, which is different than Dahl and Sterner. While both appear to be unsignificant, seasonal fluctuations are better captured by a set of seasonal dummy variables than by temperature in the stock equation. Results for the selected specifications are reported in Sections 5.3 and 5.4 of Chapter 5. Discussions in the preceding sections of the present chapter also relate to this model specification.

Model specifications for other equations were selected in a similar fashion, based on Bayesian estimates. Section 5.3 has also given a list of the explanatory variables that were tested in each alternative formulation. As different specifications lead to very similar results, in terms of predictions, we do not go through the RMSE comparison stage for other equations. Instead, generalizing the conclusions of the previous investigation, we adopted parsimonious specifications for these equations as well.

The two following sections analyze the estimation results for the selected model from a different perspective. Namely, they address the reliability and stability issues first, by asking if the estimation process converged, and then by questioning the pertinence of the hypotheses underlying the estimation and prediction technique, that is the prior choice.

6.5. Convergence

Geweke (1995) gives sufficient conditions for convergence of the Gibbs sampler. Although these were shown to hold for some of the basic estimation techniques used in the present analysis, this is not necessarily the case for more general MCMC iterative methods involving a data augmentation stage. Providing a rigorous mathematical proof that these conditions hold in our most general setting falls behind the scope of this thesis. Nevertheless, some experimental checks are performed on the MCMC draws in order to assess convergence issues.

Firstly, graphs of parameter values calculated at each iteration indicate that these seem to stabilize relatively rapidly. The same holds for the augmented values of the three dependent variables. Secondly, increasing both the total number of iterations performed (from 5000 to 10000) and the burn-in period (from 1000 to 5000) has little effect on the estimation and prediction results, apart from improving their precision. Finally, the choice of the parameter starting values has no impact on the results either. For instance, we reran the Gibbs sampling algorithms setting γ=0 and Σ=I as initial values for the GVAR model parameters and found very similar results. These all tend to indicate that convergence is achieved relatively quickly.

Assuming that convergence occurred, what remains to be checked about our estimation results are the sensitivities to the underlying assumptions. These are summarized in the prior beliefs.

6.6. Robustness to the prior choice

Prior beliefs also have a potential impact on the estimation results. In most of our applications, however, varying the whole prior distribution had minor effect on predictions. Once again, estimation processes attempting to predict the average number of light trucks and vans are the most sensitive to the underlying prior assumptions. For that equation, at least, further investigation is required.

The fact that the SUR and the GVAR models yield relatively close predictions and similar trends is reassuring. Indeed, Bayesian estimation of the SUR and GVAR models relies on distinct prior distribution assumptions. This gives credibility to the more general prior hypothesized for the GVAR model. To dig further into this issue, we would have to either vary the prior hyper-parameters of the selected GVAR model or, more generally, consider a completely different prior distribution and see how these changes affect the resulting predictions.

From a theoretical standpoint, the Normal-Diffuse prior introduced by Zellner (1971) to estimate the PVAR model is one of the most general. Unlike the widely used Minnesota prior of Litterman (1980), it allows for dependencies between equations without imposing a rigid structure on the parameter set as in the Normal-Wishart prior, for instance. According to Kadiyala and Karlsson (1993, 1997), it is also preferable in practice because it gives rise to better forecasts. Features of the chosen prior have been described in more detail in Section 4.3.2 of Chapter 4.

Another possible prior choice, which shares the same desirable properties, would be the extended natural conjugate (ENC) prior introduced by Drèze and Morales (1976). As general as the Normal-Diffuse prior, the ENC prior however suffers from several practical disadvantages. Although both priors lead to posterior distributions whose moments have no analytical closed form solutions and, hence, must be evaluated numerically, applications involving the ENC prior are particularly arduous. They both require the assumption of prior independence between equations to yield tractable posterior distributions and the specification of the prior degrees of freedom to insure that the posterior moments exist. Calculation of the ENC posterior via the Gibbs sampler can be very time consuming, especially for large models. Furthermore, the estimation results are difficult to interpret because of the reparameterization of the system. For all these reasons, estimation using the ENC prior has not been attempted.

The hyper-parameters have been set to orders of magnitude frequently encountered in the literature. They are close to the values that Litterman (1986) and Doan and al. (1984) found to work well for the Minnesota prior. In order to parameterize their priors, Kadiyala and Karlsson conduct a series of experiments in which part of the data is set aside for calibration, and the models are fitted to the remainder of the sample. Hyper-parameters are set to values minimizing the mean squared error forecasts based on the former sample. The remainder of the data is used for further investigations on estimation and forecasting performance. We cannot consider a similar approach to prior parameterization given that we need to use the whole, already modest, sample to achieve our prediction goal. To the best of our knowledge, all empirical applications of Bayesian VAR estimation techniques use hyper-parameters of the same order of magnitude as those postulated in the present study.

Kadiyala and Karlsson's experiments have shown that posterior distributions induced by the Normal-diffuse and ENC priors are more sensitive to the choice of the scale factor on the variance of own lags, π_1, than that of other dependent variables' lags, π_2 (π_3, the hyper-parameter of the variance of exogenous explanatory variables, was kept constant to a large value, reflecting prior ignorance about coefficients associated with these variables). A relevant sensitivity analysis would thus begin by varying the value of π_1, keeping both π_2 and π_3 fixed, to see how it impacts on predictions. Furthermore, our estimation results show that, among the whole lag structure, only parameters on first own-lags are significant. Others are small and not significant at conventional levels. Since π_1 determines the precision of parameters on own lags, while π_2 relates to cross effects, only changes in the value of π_1 have a potential impact on prediction results.

We have performed a series of experiments varying values of π_1 within a range of 0.05 (relatively precise prior variance) to 100 (more uncertain prior variance). As shown in Figure D.7, predictions of the average number of light trucks and vans obtained from the estimation of the GVAR(1) model do not vary much, as a result of these changes. The effect on predictions of the average distance and weighted fuel consumption rates for the same vehicle type is even weaker. We conclude that the predictions are relatively insensitive to our prior choice.

This chapter has compared all results produced in the scope of this study under several aspects in order to determine the econometric model yielding the most reliable, precise and accurate predictions of the three interest variables. The different comparisons and sensitivity analyses operated in this chapter indicate that our generalized version of Kadiyala and Karlsson's treatment for the VAR model with Normal-diffuse prior performs well in the present framework. The GVAR model allows encompassing the interrelationships among the three transport aggregate series as well as their recursive structure, in addition to exogenous economic factors. It yields reasonable, stable and relatively precise predictions. Furthermore, the estimation process happens to converge and to be relatively insensitive to the underlying assumptions and starting conditions.

Chapter 7

Conclusion

The main achievement of this study was to propose a practical, flexible and tractable econometric method for obtaining complete time series estimates from intermittent surveys. In this framework, the Bayesian approach is an advantageous alternative to classical econometric methods. Specifically, sampling-based techniques can be implemented, in collaboration with data augmentation, in order to estimate an econometric model for the survey-based variables and obtain predictions for their missing values, respectively. Such methods were applied to two Canadian surveys, the FCS and the NaPVUS, to predict transport aggregates in the private transport sector over the six-year period separating them. These exclusive sources of longitudinal information on the use of the private vehicles in Canada are crucial for prediction purpose, policy analysis, and air pollution measurement.

We were fortunate to have the first access to the newly released NaPVUS data and to acknowledge the fact that differences with the FCS cause huge discrepancies in estimates from both surveys. One important contribution of the present analysis was to propose adjustment techniques aimed at reducing such discrepancies. By accounting for distinctions in the surveys' sampling methodologies and compensating for differences in the data collection process, we produced more compatible estimates. Quarterly estimates of the three energy components, the average number of vehicles, the average distance travelled by each vehicle, and their weighted fuel consumption rates, were derived this way for cars, on one hand, and for light trucks and vans, on the other.

For each type of vehicles, an aggregate simultaneous equations model encompassing the three variables of interest was formulated. Since survey-based data were not available for the intermediary period, between the FCS and the NaPVUS, the model specification had to rely on explanatory variables drawn from external sources of information. A dynamic structure was incorporated into the model to capture the trends and seasonal variations in the quarterly time series. Potential relationships among the vehicle ownership, characteristics and use decisions were also taken into account. Since the primary interest of this empirical analysis laid in predictions, an unrestricted reduced form of the system was considered for estimation.

In order to bridge the gap between the two surveys, sampling-based techniques were generalized to include a data augmentation stage. The estimation process improved efficiency by making use of all available empirical information, while explicitly accounting for the presence of missing values in the dependent variables series. In accordance with the data augmentation principle, the missing data were considered as additional parameters that needed to be estimated. Therefore, predictions for the middle-spell variables were derived as any other Bayesian parameter estimates, endogenously and consistently with the econometric model.

We first estimated each equation separately, considering various specifications of the simple linear regression model, to examine trends and seasonal fluctuations in each series, and determine potential predictors. We also allowed for some dynamics in the single equation settings by including lagged dependent variables as regressors and autocorrelated error terms. To account for inter-relationships, as well as underlying dynamics, among the three dependent variables, simultaneous equations estimation was undertaken. In turn, we estimated seemingly unrelated regression (SUR) models and pure vector autoregressive (PVAR) formulations. By generalizing the Gibbs sampling algorithms used to estimate the two preceding models, we were able to consider a more general setting involving both models' features. The generalized vector autoregressive (GVAR) model allows for the inclusion of stochastic explanatory variables, in addition to deterministic variables and lags of each dependent variable, in each equation specification. It also accounts for possible asymmetry across the system equations.

We compared estimation and prediction results obtained for each of these model types, under alternative model specifications. Similar results were obtained in most cases, except for one of the stock equations. Precisely, predictions of the average number of light trucks and vans turned out to be highly sensitive to the hypothesized model formulation and the corresponding Bayesian estimation process. Given the irregularities showing in the observed series, this was to expect. Arguments were provided to support the fact that these were mainly caused by the recent keen interest in minivans and sports utility vehicles. The light trucks and vans stock equation was thus submitted to further tests.

A criterion was suggested to discriminate among the different model formulations and estimation techniques in order to determine the best prediction model. We argued that extrapolating predictions from both sample sides was a good way to determine how each model fitted the data and performed at completing the series. The root mean squared errors (RMSEs) of predictions obtained on small intervals surrounding the prediction period, for which data are observed, was adopted for a comparative analysis. To compute the RMSEs, each estimation process was reran and the whole series of predictions were recomputed for the enlarged missing-value zone. Within the limits of each survey's sampling period, we increased the number of missing data from one up to eight observations, acknowledging that the latter case is rather extreme, especially for the NaPVUS.

As expected, RMSEs' comparisons revealed that the more volatile NaPVUS data were harder to predict. The RMSE criterion allowed us to determine, in turn, the model type, and the model specification, that yielded the most precise and accurate predictions. Our findings indicated the importance of accounting for the relationships among vehicle ownership, characteristics and use decisions. Indeed, simultaneous equations modelling generally yielded more accurate predictions than single equation estimation. In particular, the SUR and the GVAR model specifications appeared to generate the best predictions, according to the RMSE criterion. Since the estimation results also showed the importance of accounting for serial correlation in the dependent variables series, we selected the GVAR model for prediction purposes.

Given the relatively small sample sizes involved in our applications, the Akaike and Schwartz' Bayesian information criteria favour a formulation of the GVAR involving only one-period lagged dependent variables, as opposed to more extended lag structures, for both vehicle types. Based on the RMSE criterion, a parsimonious specification involving only statistically significant stochastic exogenous variables, in addition to deterministic variables and first-order lags, was retained for each equation. This model was shown to yield the most precise and accurate predictions.

Some sensitivity analyses were performed on the selected model in order to test its reliability and robustness. Convergence issues were assessed by varying the parameter starting values as well as the number of burn-in and overall iterations performed in the Gibbs sampling algorithm. The estimation results turned out to be insensitive to these modifications. Moreover, plots of the Monte Carlo Markov Chain draws generated this way indicated that convergence is achieved relatively fast. We also varied some prior parameters of the GVAR model in order to check their impact on predictions. Overall, the results appeared to be relatively insensitive to the prior assumptions, indicating that priors do not dominate empirical information.

This study has succeeded in improving the general understanding of the private transport sector by providing Natural Resources Canada, and other governmental instances such as Environment Canada, with more accurate data on transport aggregates. These will be used for policy analysis, prediction purposes and eventually also to control air pollution. The precious sources of information provided by the two national surveys on the use of private vehicles could be further exploited to produce complete time series on other relevant variables. More disaggregated levels of the three transport aggregates are also achievable.[1]

The Bayesian approach, particularly its sampling-based techniques, has a promising future in practical applications such as this one. According to the evolving public interests, fluctuating government budget constraints and the increasing costs incumbent to undertaking surveys, more empirical information is likely to be collected on an intermittent basis, in the future. Data augmentation proved to be a powerful and tractable tool to generate complete time series in this framework, even when the proportion of missing information is relatively large, compared to the overall sample size.

The iterative algorithm developed in the context of our analysis to estimate a GVAR model appears to be privileged, at least for prediction purposes. Further investigation would be required to evaluate its performance in estimating individual structural parameters. With this respect, more general simultaneous equations models involving current-period dependent variables as regressors could also be formulated. Following Kadiyala and Karlsson (1996), other sets of priors could be considered in the GVAR framework as well. Nonlinear models could also be considered with additional computational difficulties. An interesting extension of the method would consider a model involving only survey-based data, and hence contemporaneous missing information on both sides of the equations.

The estimation and prediction techniques developed in the framework of this practical application can be seen as general solutions to the problem of missing information. They can be applied to analyze and complete other types of datasets involving missing data. In particular, they can be used to impute values to unobserved items in order to correct for the non-response or attrition bias in disaggregate cross-

[1] In the context of this empirical study, estimates by vehicle age category were also produced.

sectional surveys or panel data. Of course, they can also be employed to complete any other type of macroeconomic time series. Finally, they could also be generalized to analyze latent variables in discrete choice models.

Bibliography

Journals

1. Ben-Akiva, M. E., Manski, C. F., and Sherman, L.: A behavioural-approach model to modelling household motor-vehicle ownership and applications to aggregate policy analysis. Environment and Planning A. 13. 399 (1981)
2. Botton, K. L. and Fowkes, A. S.: An evaluation of car ownership in forecasting techniques. International Journal of Transport Economics. IV (2). 115 (1977)
3. CAA Quebec: La location fout le camp?. Touring. Winter, 65 (2000)
4. Campbell, J. Y. and Mankiw, N. G.: Are output fluctuations transitory? Quarterly Journal of Economics. 102. 857 (1987)
5. Carlin, B. P. and Polson, N. G.: Inference for nonconjugate Bayesian models using the Gibbs sampler. Canadian Journal of Statistics. 19 (4) 399 (1991)
6. Chang, M. F. et al.: Gasoline Consumption in Urban Traffic. Transportation Research Record. 599. 25 (1976)
7. Chib, S.: Bayes regression with autocorrelation errors, A Gibbs sampling approach. Journal of Econometrics. 58. 275 (1993)
8. Chib, S. and Greenberg, E.: Bayes inference in regression models with ARMA(p,q) errors. Journal of Econometrics. 64. 183 (1994)
9. Chib, S. and Greenberg, E.: Understanding the Metropolis-Hastings Algorithm. The American Statistician. 49. 327 (1995)
10. Dahl, C.: Gasoline Demand Survey. The Energy Journal. 7 (1), 67 (1986)
11. Dahl, C. and Sterner, T.: Analysing gasoline demand elasticities: a survey. Energy Economics. 13 (3), 203 (1991)

12. Deccicco, J, and Ross M.: Recent Advances in Automotive Technology and the Cost-Effectiveness of Fuel, Economy Improvements. Transportation Resources and Development. 1, 77 (1996)

13. De Jong, G. C.: An indirect utility model of car ownership and private car use. European Economic Review. 34, 971 (1990)

14. Doan, T., Litterman, R. and Sims, C.: Forecasting and conditional projection using realistic prior distributions (with discussion). Econometric Reviews. 3. 1 (1984)

15. Drèze, J. H. and Morales, J.-A.: Bayesian Full Information Analysis of Simultaneous Equations. Journal of the American Statistical Association. 71 (356). 919 (1976)

16. Duan, N., Manning, W. G., Morris, C. N. and Newhouse, J. P.: Choosing between sample-selection model and the multi-part model. Journal of Business and Economic Statistics. 2 (3), 283 (1984)

17. Eccleston, B. H. and Hurn, R. W.: Ambiant Temperature and Trip Length – Influence on Automobile Fuel Economy and Emissions. SAE Progress in Technology Series. 18. 55 (1979)

18. Fwa, T. F. and Ang, B. W.: Estimating Automobile Fuel Consumption in Urban Traffic. Transportation Research Record. 1366. 3 (1992)

19. Gallez, Caroline: Identifying the long term dynamics of car ownership: a demographic approach, Transportation Reviews, 14 (1), 83 (1994)

20. Geisser, S.: Bayesian estimation in multivariate analysis. Annals of Mathematical Statistics. 36. 150 (1965)

21. Gelfand, A. E. and Smith, A. F. M.: Sampling-Based Approaches to Calculating Marginal Densities. Journal of the American Statistical Association. 85 (410). 398 (1990)

22. Gelfand, A. E. et al.: Illustration of Bayesian inference in normal data models using Gibbs sampling. Journal of the American Statistical Association. 85. 972 (1990)

23. Geman, D. and Geman, S.: Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Transaction on Pattern Analysis and Machine Intelligence. 6. 721 (1984)

24. Geweke, J.: Antithetic Acceleration of Monte Carlo Integration in Bayesian Inference. Journal of Econometrics. 38. 73 (1988)

25. Geweke, J.: Bayesian Inference in Econometric Models using Monte Carlo Integration. Econometrica. 57. 1317 (1989)

26. Gordon, S. and Bélanger, G.: Échantillonnage de Gibbs et autres applications économétriques des chaînes markoviennes. L'actualité économique, Revue d'analyse économique. 72 (1). 27 (1996)

27. Hastings, W. K.: Monte Carlo Sampling Methods Using Markov Chains and their Applications. Biometrika. 57. 97 (1970)

28. Heckman, J.: Dummy Endogenous Variables in a Simultaneous Equation System. Econometrica. 46 (1978)

29. Hensher, D. A.: Dimensions of Automobile Demand: an Overview of an Australian Research Project. Environment and Planning. 18A. 1339-1374 (1986)

30. Hensher, D. A. and Milthorpe, F. W.: Selectivity Correction in Discrete-Continuous Choice Analysis: with Empirical Evidence for Vehicle Choice and Use. Regional Science and Urban Economics. 17. 123 (1987)

31. Jørgensen, F. and Wentzel-Larsen, T.: Forecasting Car Holding, Scrappage and New Car Purchase in Norway. Journal of Transport Economics and Policy, 24 (2). 139 (1990)

32. Kadiyala, K. R. and Karlsson, S.: Forecasting with Generalized Bayesian Vector Autoregressions. Journal of Forecasting. 12. 365 (1993)

33. Kadiyala, K. R. and Karlsson, S.: Numerical Methods for Estimation and Inference in Bayesian VAR-Models. Journal of Applied Econometrics. 12. 99 (1997)

34. Kloek, T. and van Dijk, H. K.: Bayesian Estimates of Equation System Parameters: An Application of Integration by Monte Carlo. Econometrica. 46 (1). 1 (1978)

35. Litterman, R. B.: Forecasting With Bayesian Vector Autoregressions – Five Years of Experience. Journal of Business & Economic Statistics. 4 (1). 25 (1986)

36. Mannering F. and Winston, C.: A dynamic empirical analysis of household vehicle ownership and utilization. Rand Journal of Economics. 16 (2). 215 (1985)

37. Metropolis, N. et al.: Equations of State Calculations by Fast Computing Machines. Journal of Chemical Physics. 21. 1087 (1953)

38. Natural Resources Canada, Office of Energy Efficiency: The OEE Bulletin. 1, 1 (1998)

39. Nelson, C. R. and Plosser, C. I.: Trends and random walks in macroeconomic time series: some evidence and implications. Journal of Monetary Economics. 10. 139 (1982)

40. Nerlove, M.: A note on long-run automobile demand. Journal of Marketing. 22. 57 (1957)

41. Pendyala, Ram M., Kostyniuk, Lidia P. and Goulias, Konstadinos, G.:

42. Polk New Registration: Automotive News (1994-1997)

43. Prevedouros, P. D. and Schoffer, J. L.: Factors Affecting Automobile Ownership and Use. Transportation Research Record. 1364. 152 (1992)

44. Purvis, Charles L.: Using 1990 Census Public Use Microdata Sample To Estimate Demographic and Automobile Ownership Models. Transportation Research Record 1443. 21 (1994)

45. Quebec Minister of Environment: Les changements climatiques. L'Actualité. Advertising supplement, 1 (1999)

46. Redsell, M., Lucas, G. G. and Ashford, N. J.: Factors affecting fuel consumption. Proceedings of the Institution of Mechanical Engineers. 207 (1993)

47. Reza, A. M. and Spiro, H. H.: The Demand for Passenger Car Transport Services and for Gasoline. Journal of Transport Economics and Policy. 13 (2). 304-319 (1979)

48. Roberts, G. O. and Smith, A. E. M.: Simple Conditions for the Convergence of the Gibbs Sampler and Metropolis-Hastings Algorithms. Stochastic Processes and their Applications. 49. 207 (1994)

49. Rubin, D. B.: Comment on "The Calculation of Posterior Distributions by Data Augmentation" by M. A. Tanner and W. H. Wong. Journal of the American Statistical Association. 82. 543 (1987)

50. Rubin, D. B.: Using the SIR Algorithm to Simulate Posterior Distributions. In: Bayesian Statistics (3^{d} ed.). Bernardo, J. M. and DeGroot, M. H., Lindley, D. V. and Smith, A. F. M., Oxford University Press. 395 (1988)

51. Schimek, Paul: Gasoline and Travel Demand Models Using Time Series and Cross-Section Data from United States. Transportation Research Record. 1558. 83 (1996)

52. Schipper, L. et al.: Energy use in passenger transport in OECD countries: changes since 1970. Transportation. 19 (1). 25 (1992)

53. Schwartz, G.: Estimating the dimension of a model. Annals of Statistics. 6. 461 (1978)

54. Smith, A. F. M. and Roberts, G. O.: Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. Journal of the Royal Statistical Society. B55. 3 (1993)

55. Stock, J. H. and Watson, M. W.: Variable trends in economic time series. Journal of Economic Perspective. 2. 147 (1988)

56. Tanner, J. C.: Long term forecasting of vehicle ownership and road traffic. Journal of the Royal Statistical Society. 141A (1), 14 (1978)

57. Tanner, M. A. and Wong, W. H.: The Calculation of Posterior Distributions by Data Augmentation. Journal of the American Statistical Association. 82. 528 (1987)

58. Tiao, G.C. and Zellner, A.: On the Bayesian estimation of multivariate regression. Journal of the Royal Statistical Society. B26. 389 (1964)

59. Tierney, L.: Markov Chains for Exploring Posterior Distributions. Annals of Statistics. 22 (3). 1701 (1994)

60. Transport Canada: Fuel Consumption Guide. Road Security Division. Annual Edition (1980-1999)

61. Wheaton, W. C.: The long-run structure of transportation and gasoline demand. The Bell Journal of Economics. 13 (2). 439 (1982)

Books

1. Blanchard, O. J. and Fisher S.: Lectures on Macroeconomics. MIT Press (1989)

2. Bratley, P., Fox, B. L., Schrage, L. E.: A Guide to Simulation (2nd ed.). Springler-Verlag (1983)

3. Carlin, B. P. and Thomas, A. L.: Bayes and Empirical Bayes Methods for Data Analysis (2nd ed.). Chapman & Hall (2000).

4. Chow, G.: Demand for Automobiles in the United States. North-Holland (1957)

5. Deaton, A. and Muelbauer, J.: Economics and Consumer Behaviour. Cambridge University Press (1980)

6. DeGroot, M. H.: Probability and Statistics (2nd ed.). Addison-Wesley (1986)

7. Grilliches, Z. and Intrilligator, M. D.: Handbook of Econometrics. Vol 1. North-Holland (1983)

8. Hammersley, J. M. and Handscomb, D. C.: Monte Carlo Methods. Methuen (1964)

9. Jeffreys, H.: Theory of Probability. Oxford University Press (1961)

10. Judge, G. G. et al.: The Theory and Practice of Econometrics (2nd ed.). J. Wiley and Sons, Inc. (1985)

11. Leamer, E. E.: Specification Searches. Wiley, J. and Sons (1978)

12. Mittlehammer, R. C., Judge, G. G. and Miller, D. J.: Econometric Foundations. Cambridge University Press (2000)

13. Organization for Economic Cooperation and Development: Automobile Fuel Consumption in Actual Traffic Conditions. Paris, France (1982)

14. Organization for Economic Cooperation and Development: Energy Savings and Road Traffic Management. Paris, France (1985)

15. Poirier, D.: Intermediate Statistics and Econometrics: A Comparative Approach. MIT Press (1995)

16. Smith, R. P.: Consumer Demand for Cars in the USA. Cambridge University Press (1975)

17. Stiglitz and Boadway: Principles of Macroeconomics and the Canadian Economy, (2nd ed.) Norton & Company (1994)

18. Stokey, N. L., Lucas, R. E. and Prescott, E. C.: Recursive Methods in Economic Dynamics. Harvard University Press (1989)

19. Stopher, P. R. and Meyburg, A. H.: Urban Transportation and Planning. Heath and Co. (1975)

20. Train, K.: Qualitative Choice Analysis: Theory, Econometrics and an Application to Automobile Demand. MIT Press (1986)

21. U.S. Department of Energy: Transportation Energy Data Book (1977, 1982, 1987, 1992, 1998)

22. Varian, Hal R.: Intermediate Microeconomics, A Modern Approach (2nd ed.). Norton & Company (1990)

23. Zellner, A: An Introduction to Bayesian Inference in Econometrics. Wiley & Sons (1971)

Statistical Reports

1. Natural Resources Canada: National Private Vehicle Use Survey: October-December 1994. Demand Policy and Analysis Division, Office of Energy Efficiency, Ottawa (1999)

2. Statistics Canada: Fuel Consumption Survey. Special Surveys Program. Quarterly and annual reports. Catalogues no. 53-007, 53-225, and 53-226 (1977-1988)

3. Statistics Canada: New Motor Vehicle Sales. Catalogues no. 63-007-XIB (1977-1996)

4. Statistics Canada, Private and Public Investments in Canada, Perspective 1995. Catalogue no. 61-205 (1994)

5. Statistics Canada: Quarterly Report on Energy Supply-Demand in Canada. Catalogues no. 57-003 (1976-1997)

6. Statistics Canada: Road Motor Vehicles, Registrations. Catalogues no. 53-219-XIB (1977-1996)

7. Statistics Canada: Trucking in Canada. Catalogues no. 53-222-XPB (1984-1997)

Analytical Reports

1. Bonin, Sylvie and Bernard, Marie-Christine: Quarterly fuel consumption estimates for private use vehicles in Canada, 1988-1993. Automobile Mobility Data Compendium. Final report. Document no. N96-03fa prepared for the Office of Energy Efficiency of Natural Resources Canada. Laval University, Sainte Foy (1996)

2. Bonin, Sylvie and Boucher, Nathalie: Imputation des carnets incomplets de l'Enquête nationale sur l'utilisation des véhicules privés. Automobile Mobility Data Compendium. Final report. Document no. N97-03f prepared for the Office of Energy Efficiency of Natural Resources Canada as part of the 1997-1998 work plan. Laval University, Sainte-Foy (1997)

3. Bonin, Sylvie: Méthodologie et plan d'analyse proposés pour l'imputation des données des carnets d'achats de carburant incomplets de l'ENUVeP. Automobile Mobility Data Compendium. Outline report. Document no. N99-17o prepared for the Office of Energy Efficiency of Natural Resources Canada as part of the 1990-2000 work plan. Laval University, Sainte-Foy (1999)

4. Boucher, Nathalie: Distinctions entre les nombres de véhicules à usage privé et commercial au Canada, 1980-1996. Automobile Mobility Data Compendium. Final report. Document prepared for the Office of Energy Efficiency of Natural Resources Canada as part of the 1997-1998 work plan. Laval University, Sainte-Foy (1998a)

5. Boucher, Nathalie: Étude des facteurs de marché influençant la consommation d'énergie dans le secteur du transport privé de passagers : 1990-1995. Automobile Mobility Data Compendium. Final report. Document no. N97-05f prepared for the Office of Energy Efficiency of Natural Resources Canada as part of the 1997-1998 work plan. Laval University, Sainte-Foy (1998b)

6. Boucher, Nathalie: Méthodologie proposée pour l'estimation de données nationales sur l'utilisation des véhicules privés au Canada, 1980-1996. Automobile Mobility Data Compendium. Final report. Reference document. Document prepared for the Office of Energy Efficiency of Natural Resources Canada. Laval University, Sainte-Foy (1998c)

7. Boucher, Nathalie: Merger of data from National Private Vehicle Use Survey and Manufacturer's Laboratory-tested Fuel Consumption Rates. Automobile Mobility Data Compendium. Final report. Document no. N98-15fa prepared for the Office of Energy Efficiency of Natural Resources Canada. Laval University, Sainte-Foy (1999)

8. Boucher, Nathalie: Prediction of National Data on Private Vehicle Use in Canada, 1980-1996. Automobile Mobility Data Compendium. Interim report. Document no. N98-19ia prepared for the Office of Energy Efficiency of Natural Resources Canada. Laval University, Sainte-Foy (2000)

9. Boucher, Nathalie and Bonin, Sylvie: Seasonality and Extrapolation of NaPVUS Data. Part 1: Complete National Data Series for the Period 1980-1996. Automobile Mobility Data Compendium. Interim report. Document no. N99-20f1a. prepared for the Office of Energy Efficiency of Natural Resources Canada. Laval University, Sainte-Foy (2000)

10. Environment Canada: Technical Supplement to the Environmental Indicators on Canadian Passenger Transportation. Ottawa (1996)

11. Environment Canada: Trends in Canada's Greenhouse Gas Emissions 1990 to 1995. Ottawa (1997)

12. Natural Resources Canada: Energy Efficiency Trends in Canada : 1990 to 1995. Demand Policy and Analysis Division, Office of Energy Efficiency, Ottawa (1997)

13. Quebec Minister of Energy and Natural Resources: Analyse de l'évolution de l'efficacité énergétique en transport. Annual report on the state of efficiency in Quebec. Chapter 4, Quebec, 55 (1994)

14. Royal Commission on the National Passenger Transportation Division : Final Report of the Royal Commission, Vol. 2 et 4, Chapter 7, Ottawa (1992)

15. Senneville, Annie and Bonin, Sylvie : Indicateurs pouvant expliquer l'augmentation des distances moyennes parcourues par les véhicules de promenade au Canada, Automobile Mobility Data Compendium. Final report. Document no. N96-13f prepared for the Office of Energy Efficiency of Natural Resources Canada as part of the 1996-1997 work plan (1997)

News Releases

1. Natural Resources Canada: McLellan Announces Results of Survey on Vehicle efficiency and use in Canada, Internet web site : http://www.NRCan.gc.ca/css/imb/hqlib/9699f.htm (1996)

Working Papers

1. Ayres, R. et al.: Automobile forecasting models. WP IRT-446-R. International Research and Technology Corporation. Arlington, Virginia (1976)
2. Mellman, R.: A critical analysis of automobile demand studies. Report WP 210-U2-84. Transportation Systems Center. Cambridge, Mass. (1975)
3. Transportation Table on the National Climate Change Process: Foundation Paper on Climate Change. http://www.pncc.ca (1998)

Conference papers

1. Davis, P. and Mogridge, M. J. H.: Will car ownership saturate? Some simple models of car ownership growth and examination of recent trends in OECD countries. Paper to PTRC 1976 Annual Meeting. London. Published by PTRC Education and Research Services Ltd. On behalf of Planning and Transport Research and Computation (International) Co. Ltd. (1976)

2. Whorf, R. P.: Models of automobile ownership. Proceedings of the International Conference on Transportation Research. Bruges, Belgium. Bruges: College of Europe in conjunction with Transportation Research Forum (1973)

Mimeo

1. Litterman, R. B.: A Bayesian procedure for forecasting with vector autoregressions. Mimeo. Massachusetts Institute of Technology (1980)

Appendix A

This appendix completes the information provided in Chapter 3 on survey-based data. We first present tables comparing the sample sizes of both surveys, on a quarterly basis. The adjustment process applied to the survey-based estimates is then described in further detail, when necessary.[1] Since accurate NaPVUS estimates are required for adjusting FCS estimates, the former are considered first.

A.1. Comparison of sample sizes

Table A.1 gives total sample sizes for each quarter of the FCS. In some quarters, only cars were sampled. Note that since the samples were drawn from registration files, the vehicle type of selected vehicles is always determined, so that the total of sampled cars and light trucks and vans perfectly corresponds to the total number of sampled vehicles.

Global sample sizes for each quarter of the NaPVUS are given in Table A.2. Since there were two parts to the survey, there has been double-sampling. The number of households sampled to answer the first stage telephone interview is reported in the third column. From that sample, a sub-sample of households having a private-use vehicle at their disposal was drawn to fill out the fuel purchase diary in the survey second stage. The numbers of selected vehicles sampled for that purpose are reported in subsequent columns. Note that because of some non-responses to the question regarding the type of the selected vehicle, the total number of selected vehicles may exceed the sum of sampled cars and light trucks and vans.

[1] For an extensive treatment of the adjustment process and its impact on raw survey-based estimates, refer to Boucher (2000). In the latter report, the estimates are further disaggregated by vehicle age category for each vehicle type.

Table A.1: Total sample sizes for each quarter of the FCS

Year	Quarter	Cars	Light trucks/vans	Total
1979	4	1 517	n. s.	1 517
1980	1	1 492	n. s.	1 492
1980	2	1 693	n. s.	1 693
1980	3	1 741	n. s.	1 741
1980	4	1 923	n. s.	1 923
1981	1	2 052	n. s.	2 052
1981	2	1 932	n. s.	1 932
1981	3	1 279	n. s.	1 279
1981	4	1 801	919	2 720
1982	1	1 798	714	2 512
1982	2	1 911	1 217	3 128
1982	3	2 004	1 322	3 326
1982	4	1 933	1 220	3 153
1983	1	2 000	1 276	3 276
1983	2	1 786	1 024	2 810
1983	3	2 024	1 380	3 404
1983	4	2 163	1 494	3 657
1984	1	2 121	1 487	3 608
1984	2	2 140	1 511	3 651
1984	3	2 041	1 372	3 413
1984	4	2 048	1 463	3 511
1985	1	2 031	1 446	3 477
1985	2	2 049	1 361	3 410
1985	3	1 896	1 331	3 227
1985	4	1 995	1 397	3 392
1986	1	603	449	1 052
1986	2	603	412	1 015
1986	3	589	433	1 022
1986	4	599	410	1 009
1987	1	621	438	1 059
1987	2	589	393	982
1987	3	585	399	984
1987	4	564	401	965
1988	1	579	n. s.	579
1988	2	533	n. s.	533
1988	3	474	n. s.	474
1988	4	562	n. s.	562
Means of sample sizes		1 467	1 011*	2 551*

n. s.: this particular class of vehicles was not sampled during the corresponding quarter

*: means of sample sizes are calculated on the basis of quarters during which both classes of vehicles were sampled only.

Table A.2: Total sample sizes for each quarter of the NaPVUS

Year	Quarter	Households	Cars	Light trucks/vans	Vehicles*
1994	4	5 124	1 210	497	1 707
1995	1	2 823	748	321	1 069
1995	2	2 768	687	271	960
1995	3	2 761	679	269	949
1995	4	2 797	734	341	1 078
1996	1	3 157	857	355	1 213
1996	2	3 080	861	368	1 230
1996	3	3 062	852	333	1 185
Means of sample sizes		3 197	829	344	1 174

*: the total number of sampled vehicles may not be equal to the sum of sampled cars and light trucks and vans because of non-responses regarding the vehicle type.

A.2 Adjustment to NaPVUS estimates

i) Adjustment for the non-response to certain questions

In this section, we explain the adjustment method used to bring the sum of estimates disaggregated by vehicle type into line with the total estimate. Note that this kind of adjustment is required for the NaPVUS estimates because information on some of the segmentation variables is missing. Note also that this type of adjustment is applicable to all estimates from which the variables of interest are deduced: total number of vehicles, total distance travelled, and total fuel consumption.

The adjustment is made to estimates by vehicle type in order to factor in vehicles of unspecified type. Thus, if TOTAL refers to the total estimate for all vehicles, and CAR and TRUCK denote the estimates for cars and light trucks/vans respectively, then the difference between TOTAL and SUM=CAR+TRUCK is distributed so as to reflect the initial distribution of vehicles by type on the basis of observations for vehicles of specified type. The estimates for cars will therefore be increased by a proportion P_c=CAR/SUM of the difference DIFF=TOTAL-SUM, while the residual portion of the difference, P_t=1-P_c=TRUCK/SUM, will be added to the estimates for light trucks and vans.

We will thus obtain adjusted estimates by vehicle type the sum of which will equal the total estimate for all vehicles:

$$\begin{aligned}
(\text{CAR} + P_c \times \text{DIFF}) + (\text{TRUCK} + P_t \times \text{DIFF}) &= (\text{CAR} + \text{TRUCK}) + (P_c \times \text{DIFF} + P_t \times \text{DIFF}) \\
&= \text{SUM} + (P_c + P_t) \times \text{DIFF} \\
&= \text{SUM} + [P_c + (1-P_c)] \times \text{DIFF} \\
&= \text{SUM} + \text{DIFF} \\
&= \text{SUM} + (\text{TOTAL} - \text{SUM}) \\
&= \text{TOTAL}.
\end{aligned}$$

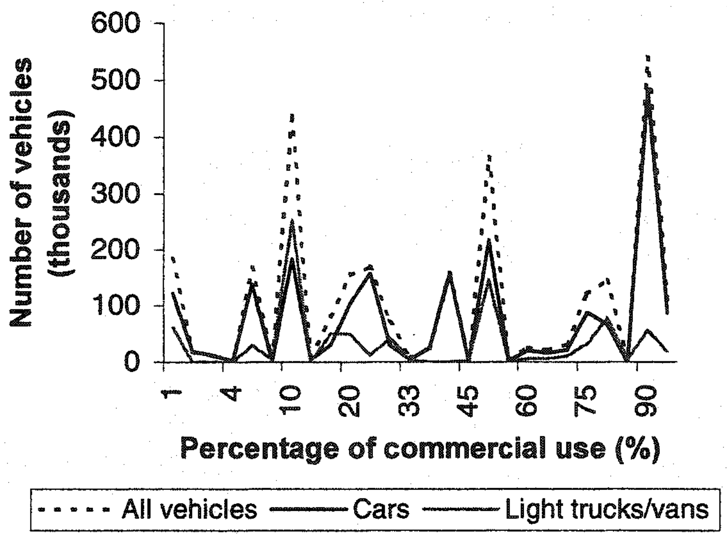
The NaPVUS estimates used in calculating the proportions in the following section underwent a similar prior adjustment. Note that the adjustments presented in the current section are applicable only to total quantities. Average quantities are then recalculated using the adjusted total quantities.

ii) Vehicles used primarily for commercial purposes

Figure A.1 provides a typical illustration of the empirical distribution of the percentage of commercial use of vehicles from the NaPVUS that are sometimes used for commercial purposes. The typical distribution has three modes around 10, 50, and 80 or 90 percent, respectively, depending on the quarter. A conservative criterion has been chosen to discriminate among vehicles according to their primary use: vehicles that are used at least 75 percent of the time for commercial purposes have been assigned to commercial vehicles.

A threshold of at least 50 percent was desired in order to keep as many vehicles as possible in the estimation basis. Furthermore, respondents whose selected vehicle was lent by their employer could have the incentive to overstate the percentage in which this vehicle was used for commercial purposes. Since the 50 percent level corresponds to a peak of the distribution, it was not an appropriate threshold. It would have been difficult to motivate assigning the exact 50 percent of commercial use to either of the private or commercial vehicle classes. The fact that there are typically only a few vehicles involving between 50 and 75 percent of commercial use motivated the choice of 75 percentage points as a threshold.

Figure A.1: Empirical distribution of the percentage of vehicle commercial use



Source : NaPVUS data for the fourth quarter of 1994.
Note: This is not a weighted distribution.

A.3 Adjustments to FCS estimates

In this section, we present the main points of the adjustments made to the FCS estimates so that they would be comparable to the NaPVUS estimates. Only total estimates of number of vehicles, distance and fuel consumption are considered here. We then use the adjusted figures to generate estimates of average distance and weighted fuel consumption rate, since these too are of significant interest to us in the context of this study.

i) Adjustment for vehicles excluded from the sampling field

The first adjustment, for vehicle categories that were excluded from the registration files of certain provinces for some quarters, is based on Statistics Canada estimates and has been explained in detail in section 3.6.1 of Chapter 3. A list of the unsurveyed categories is presented in Table A.3.

Table A.3: Categories of vehicles excluded from provincial registration files at the moment of the FCS sampling

	NWF		QUE		ONT		ALB		BC	
Year-quarter	Cars	Light trucks /vans	Cars	Light trucks /vans	Cars	Light trucks /vans	Cars	Light trucks /vans	Cars	Light trucks /vans
79-4			T							
80-1			L		T					
80-2			L							
81-4			L	L	L	L			L	L
82-1			L	L		T			L	L
82-2			L	L						
82-4	L	L			L	L				
83-1					L	L				
83-2					L	L			T	T
85-1							L	L		
85-2							L	L		

Legend : T : the totality of vehicles were absent from the registration files for the corresponding province and quarter.

L : only late model vehicles are missing from the registration files for the corresponding province and quarter.

Subsequent adjustments are based on the assumption that the vehicle distribution remained basically the same throughout the period delimited by the two surveys. The assumption may not be realistic, but in our view it is better to make the adjustments and thus make the estimates for the two surveys comparable rather than using just the raw estimates from the surveys.

Indeed, NaPVUS is the only available source of information that can give us an idea of the type of adjustments to the FCS estimates required in order to factor in the exclusion of certain vehicle categories from that survey. Furthermore, changes in vehicle distribution probably result in inverse moves in correction terms. More specifically, for the next two types of adjustments based on the NaPVUS distribution, variations in vehicle distribution over the last twenty years probably moved in opposite directions.

Vehicles that were not driven during the survey month were probably more numerous in the eighties than during the NaPVUS survey period. Therefore, a NaPVUS-based adjustment to factor in the exclusion of vehicles not driven during the FCS survey month will generate an underestimation of the number of non-driven vehicles. Indeed, the use of vehicles for private purposes has increased in recent years, compared with the eighties. Furthermore, given the economic and demographic trends, the number of new vehicles purchased or leased on a long term basis has increased over the same period.[2] As a result, the adjustment for new vehicles on the basis of the proportion of such vehicles in the NaPVUS may well overestimate the number of new vehicles in the FCS.

By way of illustration, we have distributed the NaPVUS vehicles as shown in Table A.4. Portion (1) is the number of privately owned or leased NaPVUS vehicles that are used primarily for private purposes, were driven during the survey month, and are not new. The definition of a new vehicle (current or following year vehicle) may change here, depending on the FCS quarter for which the adjustment is required (see Section 3.6.3 of Chapter 3): Portion (1) does not include the NaPVUS vehicles that are not

[2] See Boucher (1998b) for a justification of the two last assertions.

represented in registration files for the corresponding FCS survey quarter. Therefore, in their turn, the subsequent adjustments add portions (2) and (3) to the FCS estimates.

Table A.4: Distribution of the NaPVUS vehicles for the FCS adjustments

Portion (1)	Portion (2)	Portion (3)	Portion (4)
FCS vehicles covered by the first adjustment	Non-driven vehicles	New vehicles	Vehicles used primarily, but not solely, for commercial purposes

Recall that the proportions represented by each of the above portion in the NaPVUS are based on two quarters. Since the NaPVUS was conducted for two full years, the proportion for a specific quarter is based on NaPVUS observations for the two corresponding quarters. We were thus able to generate quarterly adjustments based on as many observations as possible and thereby make the results more significant. Note that, prior to such a calculation, we made an adjustment similar to those described in section 3.5.2 of Chapter 3 in order to factor in the missing values for the NaPVUS variables needed to calculate the proportions.

Total estimates by vehicle type are adjusted using the corresponding NaPVUS proportions. In other words, the FCS-based estimates are increased by the proportion of the corresponding category in the NaPVUS. Consistency is ensured by the fact that proportions by type are calculated in relation to the grand total, including non-driven vehicles. Here is a detailed description of the adjustment process, step by step.

ii) Adjustment for unused vehicles

The purpose of the second adjustment is to factor in the exclusion from the FCS of vehicles that were not driven during the survey month. Only estimates of the total number of vehicles are adjusted, because distance and fuel consumption for those vehicles show nil values. Thus the proportion of non-driven vehicles in NaPVUS is determined by calculating the ratio of portion (1) to the total of portions (1) and (2). A quarterly proportion of non-driven vehicles can thus be obtained for each vehicle type (PROPND).

In order to clarify the linkages in the adjustment process, we will use FCS^j, for j=1,...,4, to denote the FCS estimates by vehicle type derived from the j^{th} adjustment. Thus the second adjustment applies to estimates resulting from the first StatCan adjustment-based, i.e. to FCS^1, and it will generate adjusted estimates, FCS^2. Estimates by vehicle type are increased by a PROPND proportion of the adjusted total, FCS^2 :

$$FCS^2 = FCS^1 + (PROPND \times FCS^2).$$

The adjusted estimates are thus obtained with the formula :

$$FCS^2 = \frac{FCS^1}{1 - PROPND}.$$

Adjusted estimates resulting from the second adjustment equal the sum of portions (1) and (2) in Table A.4.

iii) Adjustment for new vehicles

The purpose of the third adjustment is to add new vehicle categories that were not surveyed in the FCS. The categories undergoing this adjustment vary, depending on the FCS quarter under consideration (see section 3.6.3). A correction is made to all three variables: number of vehicles, distance travelled, and fuel consumption. The proportion of the value of one of these variables to be assigned to new vehicles, PROPN, for a given vehicle type is calculated on the basis of the NaPVUS observations for the corresponding quarters by dividing portion (3), which represents the total quantity of the variable to be assigned to new vehicles excluded from the FCS for the corresponding quarter, by the sum of the same variable for portions (1), (2) and (3).

To adjust an estimate of a variable, FCS^2, resulting from the second adjustment for a given vehicle type, we add to it the missing portion for new vehicles. As in the preceding correction, the adjusted estimate takes the form:

$$FCS^3 = FCS^2 + \left(PROPN \times FCS^3\right),$$

and it can be calculated using the formula :

$$FCS^3 = \frac{FCS^2}{1 - PROPN}.$$

The last attempt to retrieve the remaining primarily commercial-use vehicles (portion (4)) from the resulting estimates was unsuccessful, for reasons explained in section 3.6.4.

Appendix B

This appendix presents the simulation results discussed in Section 5.1 of Chapter 5. The data generating process (DGP) for each series of Monte Carlo experiments is first described. Then, simulation results are tabulated and compared with real (assumed) parameter values appearing in the headline of each table. Graphs of predictions and illustrations of their dispersion and accuracy are also provided to evaluate the prediction performance of each type of model. We start by considering single equation estimation and then turn to the estimation of systems of simultaneous equations. In all cases, the appropriate sampling-based technique is applied to obtain the estimates and predictions. The starting values and prior assumptions are those described in Chapter 4, where the Bayesian approach is introduced.

1. First simulation: Simple linear regression (LIN) model

DGP: $Y=\beta_0+\beta_1X+u$, where $u \sim N(0,\sigma^2 I_T)$ with $x_t \sim NID(4,2)$, for $t=1,\ldots,100$.

Table B.1: Simulation results for the simple linear regression model

Nu. of missing observations	Prediction interval	$\beta_0=1.000$	$\beta_1=0.300$	$\sigma^2=0.200$
0	----	0.874 (0.169)	0.331 (0.038)	0.246 (0.032)
10	[46,55]	0.858 (0.187)	0.341 (0.045)	0.257 (0.036)
20	[41,60]	0.915 (0.209)	0.327 (0.050)	0.234 (0.037)
30	[36,65]	0.880 (0.220)	0.338 (0.052)	0.251 (0.041)
40	[31,70]	0.735 (0.224)	0.363 (0.051)	0.235 (0.050)
50	[26,75]	0.764 (0.211)	0.348 (0.051)	0.244 (0.049)
60	[21,80]	0.795 (0.240)	0.336 (0.054)	0.274 (0.055)
70	[16,85]	1.023 (0.366)	0.281 (0.078)	0.279 (0.078)
80	[11,90]	0.901 (0.394)	0.303 (0.091)	0.329 (0.164)
90	[6,95]	0.815 (0.579)	0.310 (0.143)	0.600 (0.280)

Legend: Numbers in parentheses correspond to standard errors on the parameter estimates.

Figure B.1: Predictions of the simple linear regression model based on simulated data

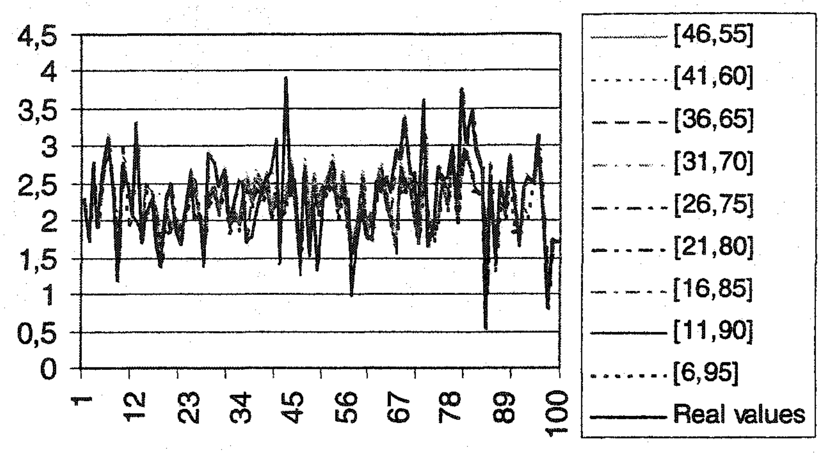


Figure B.2: Comparison of predictions of the simple linear regression model with simulated data on the [26,75] interval

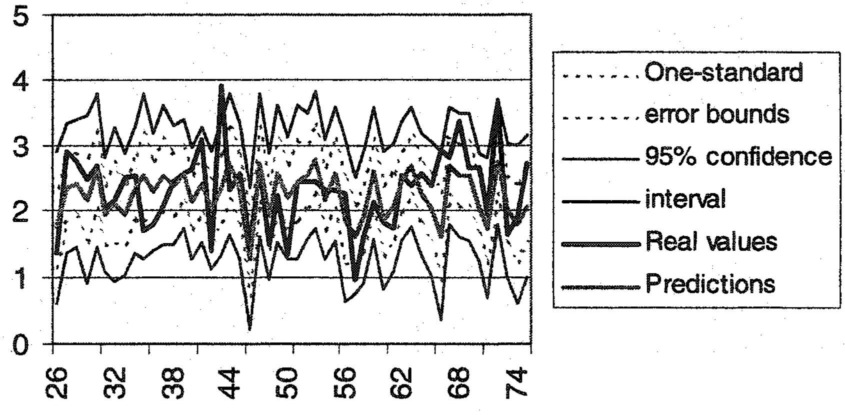
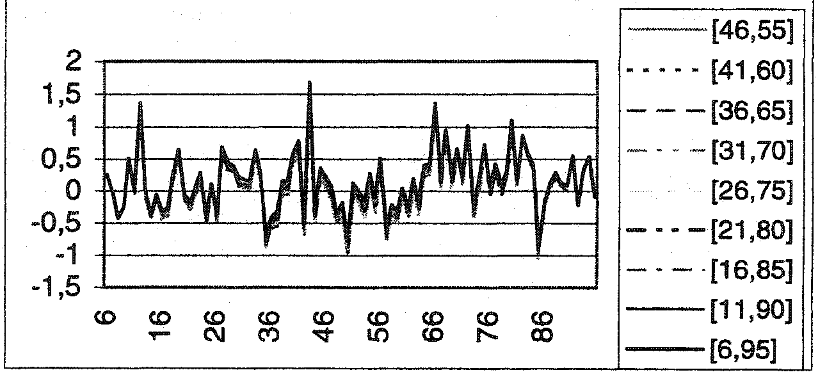


Figure B.3: Prediction errors of the simple linear regression model estimated on simulated data



2. Second simulation: Linear regression model involving a four-period lagged dependent variable (PLAG(4)) as regressor

DGP: $y_t=\beta_0+\alpha y_{t-4}+u_t$, where $u_t \sim N(0,\sigma^2 I_T)$, for t=1,...,100.

Note: We start from $y_0=0$ and skip 300 turns before keeping the simulated data.

Table B.2: Simulation results for the linear regression model including a four-period lagged dependent variable as regressor

Nu. of missing observations	Prediction interval	$\beta_0=1$	$\alpha=0.3$	$\sigma^2=0.2$
0	----	0.949 (0.138)	0.335 (0.090)	0.226 (0.039)
10	[46,55]	0.928 (0.164)	0.356 (0.117)	0.237 (0.047)
20	[41,60]	0.936 (0.165)	0.349 (0.119)	0.243 (0.045)
30	[36,65]	0.902 (0.181)	0.370 (0.132)	0.246 (0.047)
40	[31,70]	0.822 (0.216)	0.416 (0.152)	0.243 (0.055)
50	[26,75]	0.839 (0.246)	0.387 (0.162)	0.255 (0.062)
60	[21,80]	0.875 (0.234)	0.383 (0.151)	0.264 (0.073)
70	[16,85]	0.838 (0.322)	0.408 (0.200)	0.269 (0.078)
80	[11,90]	0.920 (0.346)	0.391 (0.226)	0.393 (0.202)
90	[6,95]	1.232 (0.517)	0.168 (0.297)	1.617 (1.872)

Legend: Numbers in parentheses correspond to standard errors on the parameter estimates.

Figure B.4: Predictions of the linear regression model involving a four-period lagged dependent variable based on simulated data

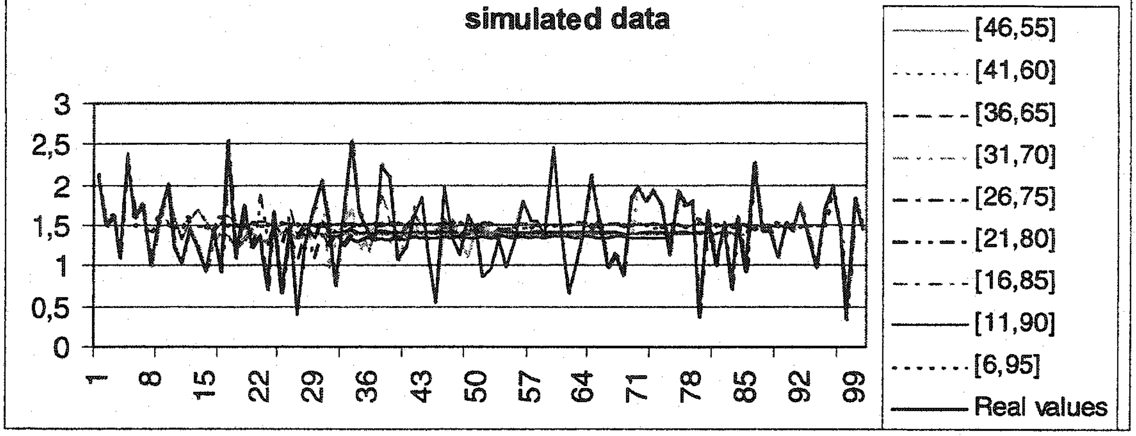


Figure B.5: Comparison of predictions of the linear regression model involving a four-period lagged dependent variable with simulated data on the [31,70] interval

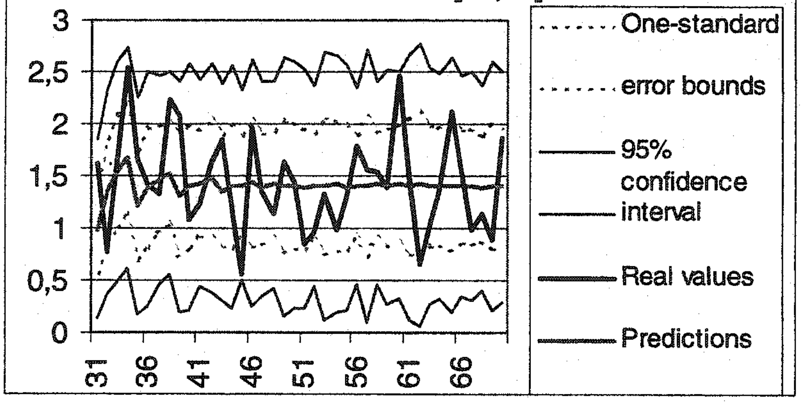
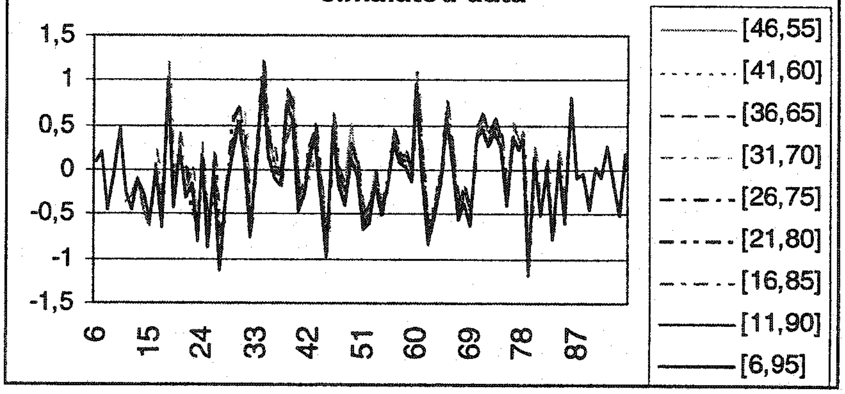


Figure B.6: Prediction errors of the linear regression model involving a four-period lagged dependent variable estimated on simulated data



3. Third simulation: Linear regression model with AR(4) errors

DGP: $y_t=\beta_0+\beta_1x_t+e_t$, where $e_t=\phi_1e_{t-1}+\phi_2e_{t-2}+\phi_3e_{t-3}+\phi_4e_{t-4}+u_t$, with u_t~NID(0,σ^2) and x_t~NID(4,2), for t=1,…,100.

Note: We start from $e_0=e_1=e_2=e_3=0$, and skip 300 turns before keeping the simulated data.

Table B.3: Simulation results for the linear regression model with AR(4) errors

Nu. of miss. obs.	Pred. interval	β_0=1.0	β_1=0.3	ϕ_1=-0.4	ϕ_2=-0.2	ϕ_3=-0.1	ϕ_4=0.5	σ^2=0.2
0	----	0.963 (0.110)	0.305 (0.025)	-0.458 (0.099)	-0.099 (0.125)	-0.017 (0.115)	0.526 (0.080)	0.255 (0.045)
10	[46,55]	0.977 (0.122)	0.300 (0.029)	-0.515 (0.118)	-0.127 (0.118)	-0.023 (0.150)	0.466 (0.101)	0.310 (0.053)
20	[41,60]	1.081 (0.194)	0.281 (0.048)	-0.537 (0.107)	-0.151 (0.128)	-0.028 (0.136)	0.441 (0.091)	0.339 (0.070)
30	[36,65]	0.994 (0.173)	0.300 (0.043)	-0.488 (0.132)	-0.126 (0.131)	-0.039 (0.160)	0.454 (0.122)	0.348 (0.101)
40	[31,70]	1.014 (0.214)	0.304 (0.048)	-0.470 (0.139)	-0.135 (0.178)	-0.058 (0.179)	0.445 (0.130)	0.358 (0.092)
50	[26,75]	1.111 (0.224)	0.281 (0.058)	-0.462 (0.159)	-0.128 (0.122)	-0.062 (0.151)	0.430 (0.154)	0.416 (0.102)
60	[21,80]	1.053 (0.286)	0.289 (0.075)	-0.429 (0.135)	-0.103 (0.143)	-0.070 (0.186)	0.414 (0.156)	0.514 (0.131)
70	[16,85]	1.042 (0.406)	0.296 (0.103)	-0.419 (0.141)	-0.105 (0.193)	-0.110 (0.204)	0.342 (0.173)	0.635 (0.164)
80	[11,90]	0.598 (0.515)	0.408 (0.124)	-0.385 (0.221)	0.217 (0.294)	0.074 (0.244)	0.240 (0.203)	0.780 (0.247)
90	[6,95]	0.943 (0.611)	0.290 (0.203)	-0.294 (0.270)	0.076 (0.304)	-0.161 (0.309)	-0.051 (0.316)	1.726 (0.769)

Legend: Numbers in parentheses correspond to standard errors on the parameter estimates.

Figure B.7: Predictions of the linear regression model with AR(4) errors based on simulated data

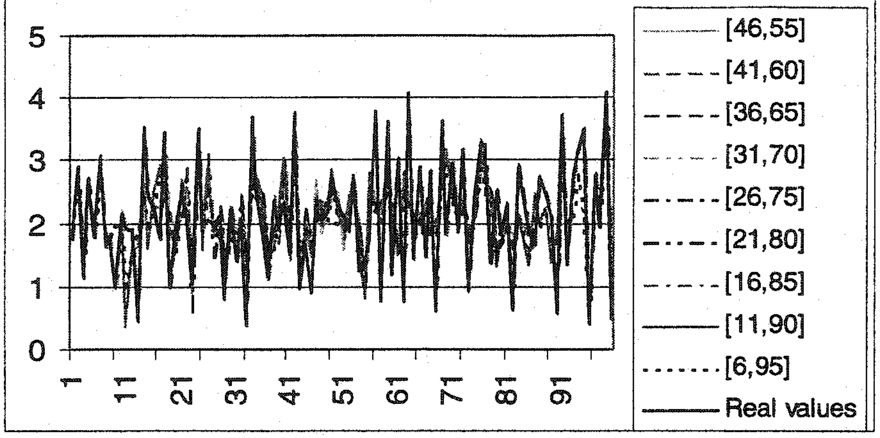


Figure B.8: Comparison of predictions of the linear regression model involving AR(4) errors with simulated data on the [26,75] interval

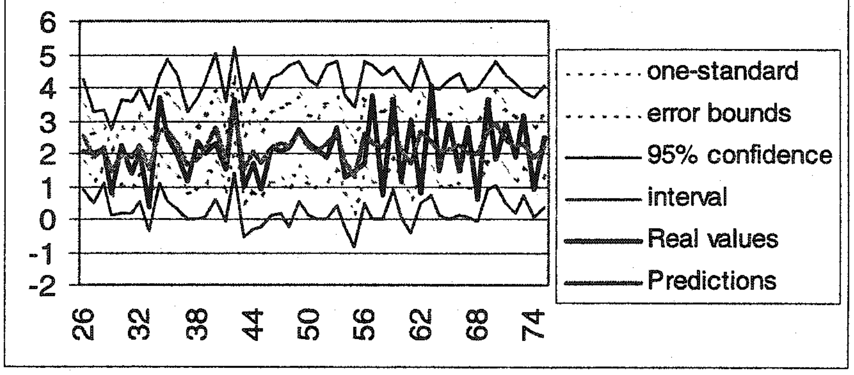
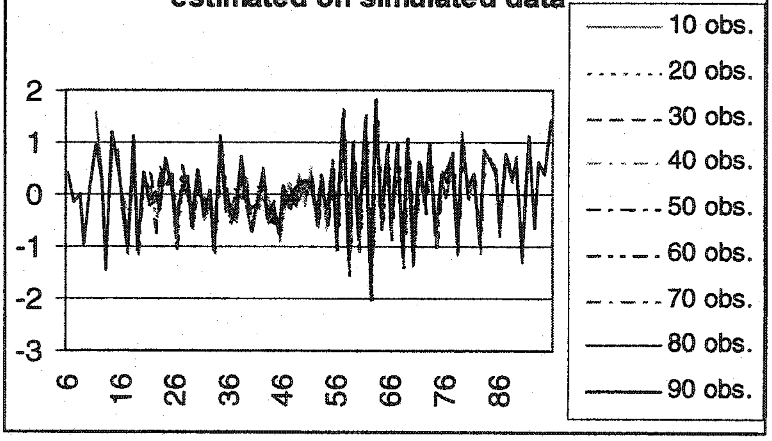


Figure B.9: Prediction errors of the linear regression model with AR(4) errors estimated on simulated data



4. Fourth simulation: Seemingly unrelated regression (SUR) model

DGP: $y_{1t}=\theta_{10}+\theta_{11}x_{1t}+\varepsilon_{1t}$

$y_{2t}=\theta_{20}+\theta_{21}x_{2t}+\varepsilon_{2t}$

where x_{1t}~NID(4,2) and x_{2t}~NID(3,1.5) are not correlated, and ε_t~NID(0,Σ), for t=1,...100.

Table B.4: Simulation results for the seemingly unrelated equations model

Nu. of miss. obs.	θ_{10}=1.0	θ_{11}=0.3	θ_{20}=2.0	θ_{21}=0.5	Σ_{11}=0.2	Σ_{12}=0.0	Σ_{22}=0.4
0	0.903 (0.128)	0.331 (0.030)	1.924 (0.204)	0.542 (0.063)	0.214 (0.028)	0.005 (0.041)	0.498 (0.069)
10	0.905 (0.173)	0.330 (0.040)	1.900 (0.156)	0.552 (0.049)	0.210 (0.031)	-0.006 (0.035)	0.468 (0.085)
20	0.812 (0.168)	0.346 (0.040)	2.008 (0.197)	0.523 (0.065)	0.209 (0.037)	-0.002 (0.044)	0.493 (0.073)
30	0.871 (0.175)	0.329 (0.039)	1.944 (0.161)	0.513 (0.051)	0.184 (0.041)	-0.015 (0.035)	0.432 (0.087)
40	0.966 (0.175)	0.304 (0.043)	1.855 (0.223)	0.518 (0.063)	0.186 (0.043)	-0.008 (0.041)	0.392 (0.070)
50	0.880 (0.172)	0.323 (0.041)	1.855 (0.194)	0.515 (0.063)	0.185 (0.031)	-0.005 (0.038)	0.401 (0.105)
60	0.935 (0.264)	0.314 (0.056)	1.845 (0.280)	0.542 (0.087)	0.206 (0.047)	-7×10^{-5} (0.043)	0.418 (0.085)
70	1.124 (0.233)	0.265 (0.055)	1.797 (0.403)	0.567 (0.124)	0.206 (0.051)	0.024 (0.059)	0.498 (0.123)
80	1.151 (0.428)	0.253 (0.096)	1.763 (0.428)	0.580 (0.125)	0.277 (0.149)	0.003 (0.101)	0.595 (0.245)
90	0.767 (0.457)	0.369 (0.118)	1.765 (0.642)	0.560 (0.243)	0.355 (0.172)	-0.004 (0.230)	0.683 (0.876)

Legend: Numbers in parentheses correspond to standard errors on the parameter estimates.

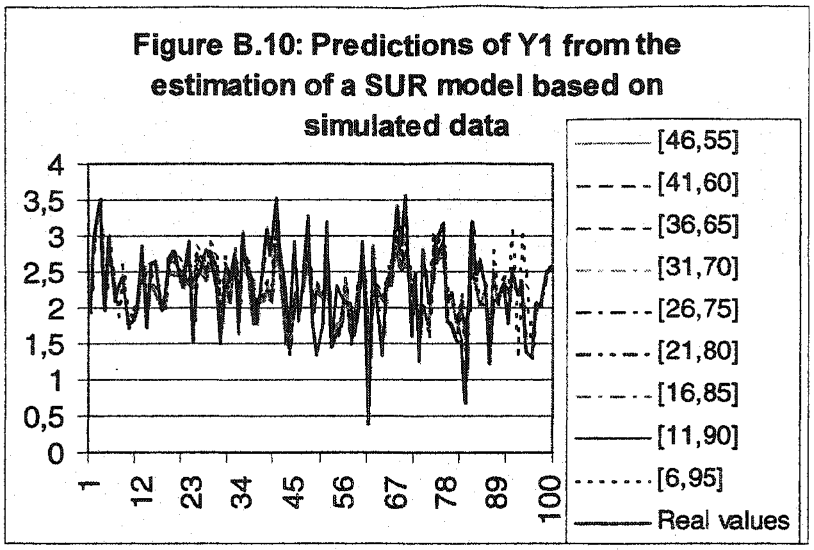


Figure B.10: Predictions of Y1 from the estimation of a SUR model based on simulated data

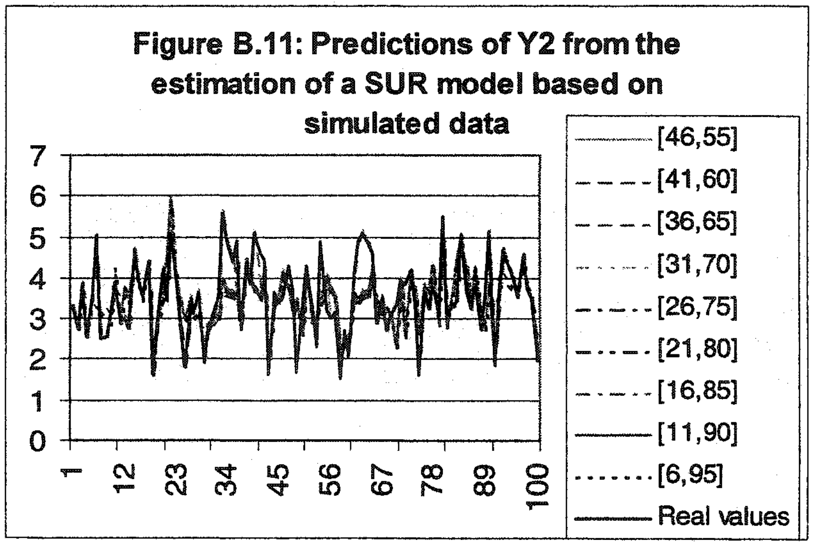


Figure B.11: Predictions of Y2 from the estimation of a SUR model based on simulated data

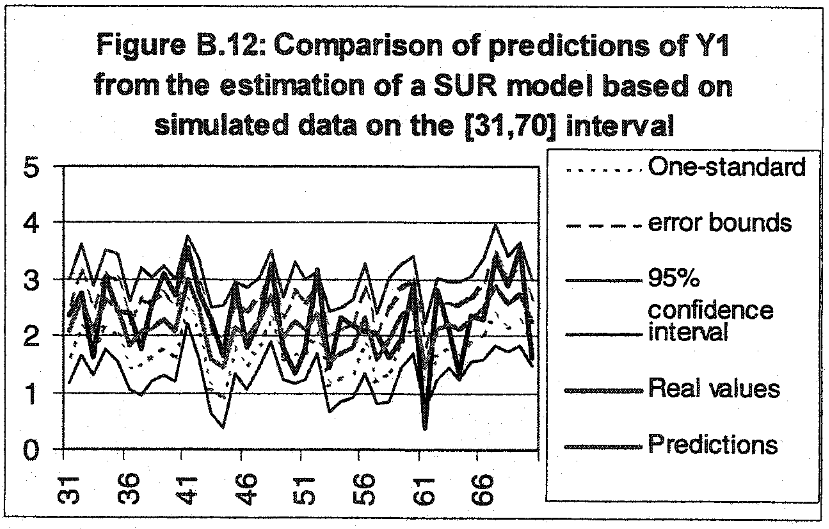


Figure B.12: Comparison of predictions of Y1 from the estimation of a SUR model based on simulated data on the [31,70] interval

Figure B.13: Comparison of predictions of Y2 from the estimation of a SUR model based on simulated data on the [31,70] interval

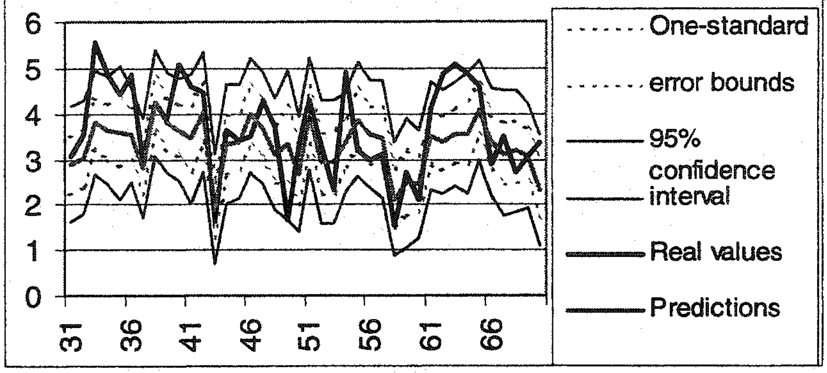


Figure B.14: Prediction errors on Y1 from the estimation of a SUR model based on simulated data

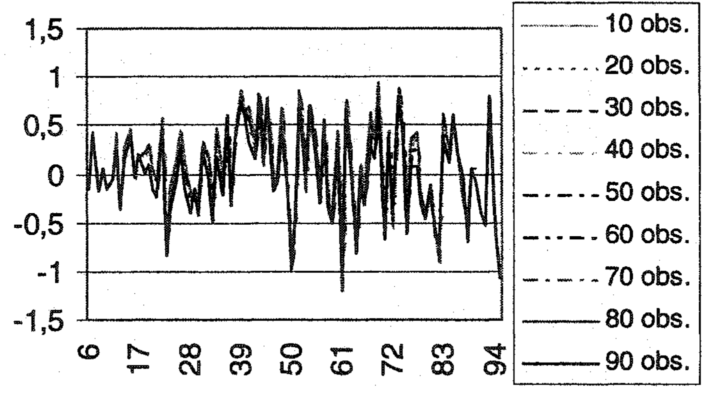
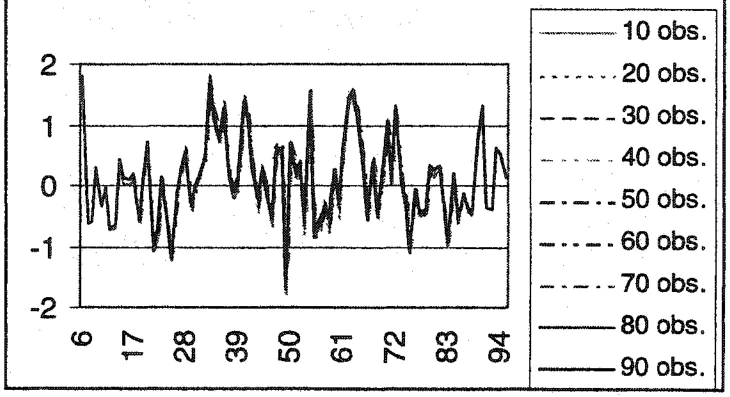


Figure B.15: Prediction errors on Y2 from the estimation of a SUR model based on simulated data



5. Fifth simulation: Pure vector autoregressive model of order 1 (PVAR(1))

DGP: $y_{1t}=\theta_{10}+\varphi_{11}y_{1t-1}+\varphi_{12}y_{2t-1}+e_{1t}$

$y_{2t}=\theta_{20}+\varphi_{21}y_{1t-1}+\varphi_{22}y_{2t-1}+e_{2t}$, where $e_t=[e_{1t}\ \ e_{2t}]\sim NID(0,\Sigma)$, for t=1,...,100.

Note: We start from $y_{10}=y_{20}=0$, and skip 300 turns before keeping the simulated data.

Table B.5: Simulation results for the pure vector autoregressive model of order 1

Missing obs.	0	10	20	30	40	50	60	70	80	90
$\theta_{10}=1.0$	1.087 (0.300)	1.122 (0.322)	1.069 (0.331)	0.977 (0.314)	0.947 (0.274)	0.885 (0.427)	0.956 (0.368)	0.998 (0.557)	0.900 (0.796)	-0.116 (21.527)
$\varphi_{11}=0.5$	0.619 (0.073)	0.619 (0.079)	0.635 (0.079)	0.658 (0.068)	0.666 (0.062)	0.699 (0.100)	0.688 (0.091)	0.677 (0.123)	0.724 (0.151)	0.733 (0.132)
$\varphi_{12}=0.4$	0.193 (0.044)	0.172 (0.045)	0.169 (0.050)	0.166 (0.050)	0.159 (0.047)	0.118 (0.040)	0.106 (0.044)	0.098 (0.053)	0.070 (0.055)	0.019 (0.075)
$\theta_{20}=3.0$	2.035 (0.323)	2.009 (0.434)	1.900 (0.429)	1.896 (0.350)	1.822 (0.475)	1.752 (0.560)	1.473 (0.585)	1.193 (0.504)	1.055 (0.465)	1.078 (4.650)
$\varphi_{21}=-0.5$	-0.277 (0.077)	-0.256 (0.095)	-0.248 (0.079)	-0.240 (0.077)	-0.232 (0.087)	-0.169 (0.098)	-0.126 (0.113)	-0.101 (0.090)	-0.064 (0.086)	-0.007 (0.056)
$\varphi_{22}=0.7$	0.674 (0.063)	0.657 (0.086)	0.682 (0.067)	0.671 (0.082)	0.684 (0.108)	0.620 (0.133)	0.687 (0.115)	0.753 (0.110)	0.772 (0.099)	0.743 (0.129)
$\Sigma_{11}=0.2$	0.225 (0.036)	0.252 (0.042)	0.259 (0.046)	0.255 (0.070)	0.282 (0.060)	0.320 (0.068)	0.253 (0.127)	0.333 (0.124)	26.000 (46.476)	3942.577 ($7\text{x}10^4$)
$\Sigma_{12}=0.0$	-0.013 (0.041)	0.002 (0.047)	-0.004 (0.050)	0.026 (0.061)	0.033 (0.071)	0.092 (0.086)	0.012 (0.112)	-0.017 (0.116)	2.559 (3.940)	-676.994 ($1\text{x}10^4$)
$\Sigma_{22}=0.4$	0.508 (0.099)	0.533 (0.082)	0.571 (0.114)	0.621 (0.110)	0.672 (0.127)	0.862 (0.239)	0.683 (0.160)	0.783 (0.222)	54.755 (53.324)	145.631 ($2\text{x}10^3$)

Legend: Numbers in parentheses correspond to standard errors on the parameter estimates.

Figure B.16: Predictions of Y1 from the estimation of a PVAR(1) model based on simulated data

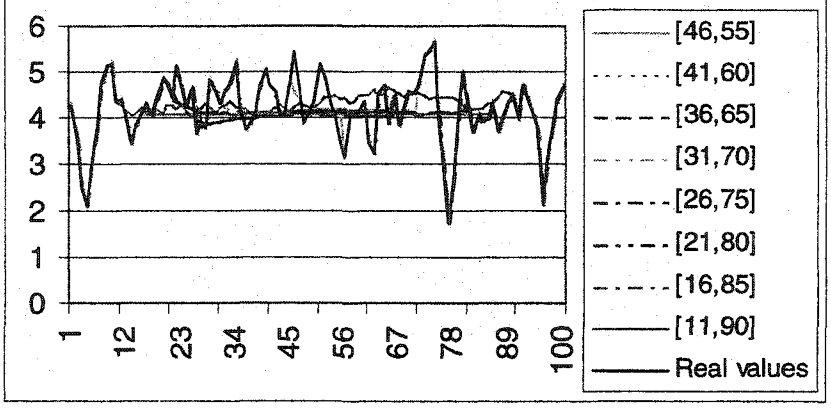


Figure B.17: Predictions of Y2 from the estimation of a PVAR(1) model based on simulated data

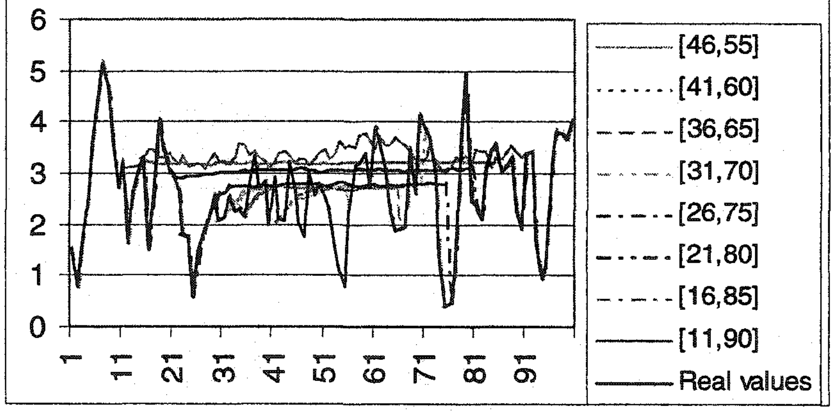


Figure B.18: Comparison of predictions of Y1 from the estimation of a PVAR(1) model based on simulated data on the [26,75] interval

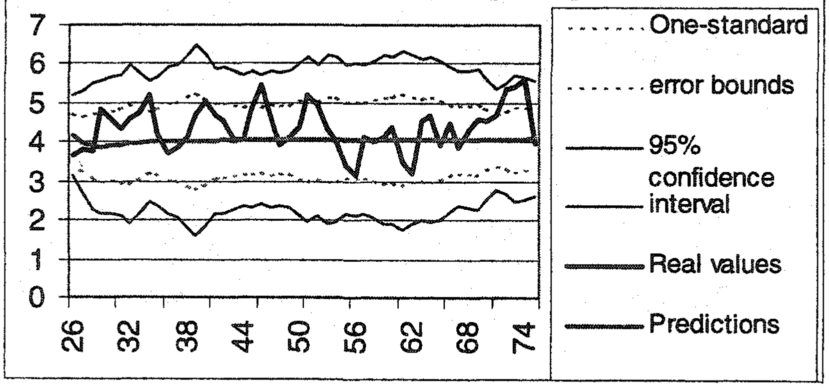


Figure B.19: Comparison of predictions of Y2 from the estimation of a PVAR(1) model based on simulated data on the [26,75] interval

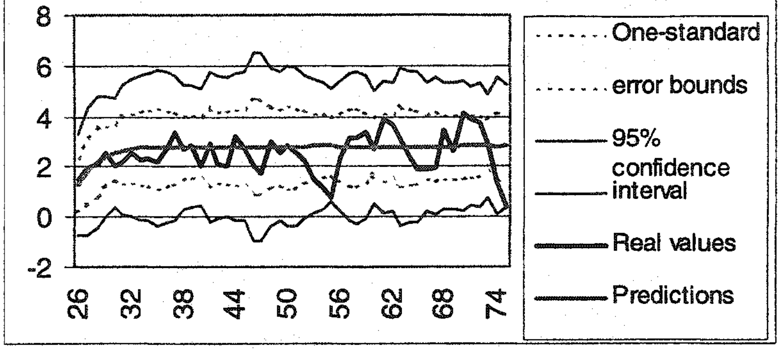


Figure B.20: Prediction errors on Y1 from the estimation of a PVAR(1) model based on simulated data

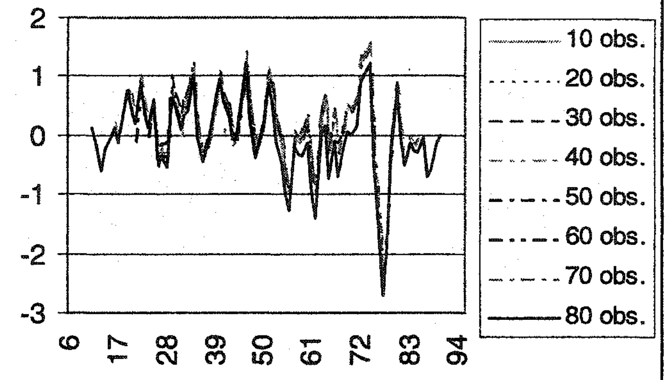
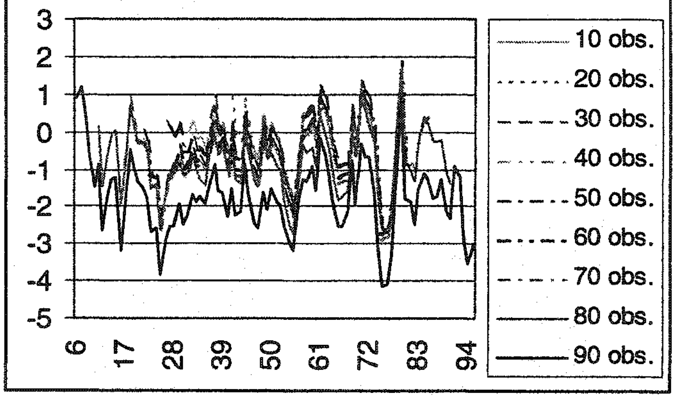


Figure B.21: Prediction errors on Y2 from the estimation of a PVAR(1) model based on simulated data



6. Sixth simulation: Generalized vector autoregressive model of order 1 (GVAR(1))

DGP: $y_{1t}=\Delta_{10}+\Delta_{11}x_{1t}+\Delta_{12}y_{1t-1}+\Delta_{13}y_{2t-1}+u_{1t}$

$y_{2t}=\Delta_{20}+\Delta_{21}x_{2t}+\Delta_{22}y_{1t-1}+\Delta_{23}y_{2t-1}+u_{2t}$

where ε_t~NID(0,Σ), T=100, and x_{1t}~NID(4,2) and x_{2t}~NID(3,1.5) are not correlated.

Note: We start from $y_{10}=y_{20}=0$, and skip 300 turns before keeping the simulated data.

Table B.6: Simulation results for the generalized vector autoregressive model of order 1

Missing obs.	0	10	20	30	40	50	60	70	80	90
Δ_{10}=1.0	0.822 (0.356)	0.890 (0.463)	0.817 (0.531)	0.545 (0.394)	0.620 (0.626)	0.415 (0.742)	0.623 (0.669)	0.878 (0.777)	0.526 (1.048)	0.842 (6.244)
Δ_{11}=0.2	0.197 (0.031)	0.178 (0.038)	0.197 (0.043)	0.200 (0.034)	0.179 (0.049)	0.183 (0.039)	0.167 (0.050)	0.198 (0.085)	0.217 (0.107)	0.138 (1.473)
Δ_{12}=0.5	0.632 (0.058)	0.644 (0.071)	0.651 (0.080)	0.691 (0.055)	0.699 (0.093)	0.725 (0.111)	0.733 (0.109)	0.680 (0.123)	0.710 (0.143)	0.711 (0.158)
Δ_{13}=0.4	0.205 (0.035)	0.183 (0.034)	0.173 (0.050)	0.179 (0.064)	0.162 (0.048)	0.158 (0.056)	0.108 (0.051)	0.065 (0.057)	0.054 (0.063)	0.012 (0.077)
Δ_{20}=3.0	1.328 (0.393)	1.359 (0.434)	1.192 (0.569)	1.312 (0.584)	0.984 (0.476)	0.777 (0.591)	0.722 (0.654)	0.562 (0.675)	0.488 (0.669)	-0.687 (4.775)
Δ_{21}=0.3	0.327 (0.066)	0.334 (0.064)	0.318 (0.091)	0.315 (0.061)	0.309 (0.068)	0.336 (0.114)	0.317 (0.083)	0.265 (0.141)	0.201 (0.173)	0.446 (1.701)
Δ_{22}=-0.5	-0.241 (0.069)	-0.243 (0.064)	-0.218 (0.082)	-0.206 (0.083)	-0.166 (0.065)	-0.105 (0.067)	-0.135 (0.084)	-0.094 (0.081)	-0.054 (0.088)	-0.004 (0.058)
Δ_{23}=0.7	0.696 (0.058)	0.672 (0.064)	0.696 (0.083)	0.630 (0.099)	0.660 (0.112)	0.578 (0.120)	0.694 (0.097)	0.712 (0.121)	0.705 (0.138)	0.683 (0.164)
Σ_{11}=0.2	0.238 (0.036)	0.260 (0.045)	0.270 (0.049)	0.253 (0.055)	0.282 (0.079)	0.306 (0.104)	1.170 (0.913)	0.344 (0.100)	1.632 (1.256)	68.643 (221.075)
Σ_{12}=0.0	-0.032 (0.034)	-0.027 (0.037)	-0.022 (0.048)	-0.017 (0.046)	-0.046 (0.071)	$-6\text{x}10^{-4}$ (0.083)	-0.097 (0.092)	-0.049 (0.102)	-0.359 (0.388)	-3.577 (22.650)
Σ_{22}=0.4	0.395 (0.062)	0.423 (0.059)	0.415 (0.077)	0.392 (0.083)	0.464 (0.113)	0.587 (0.149)	0.502 (0.121)	0.507 (0.286)	2.009 (1.769)	41.567 (158.344)

Legend: Numbers in parentheses correspond to standard errors on the parameter estimates.

Figure B.22: Predictions of Y1 from the estimation of a GVAR(1) model based on simulated data

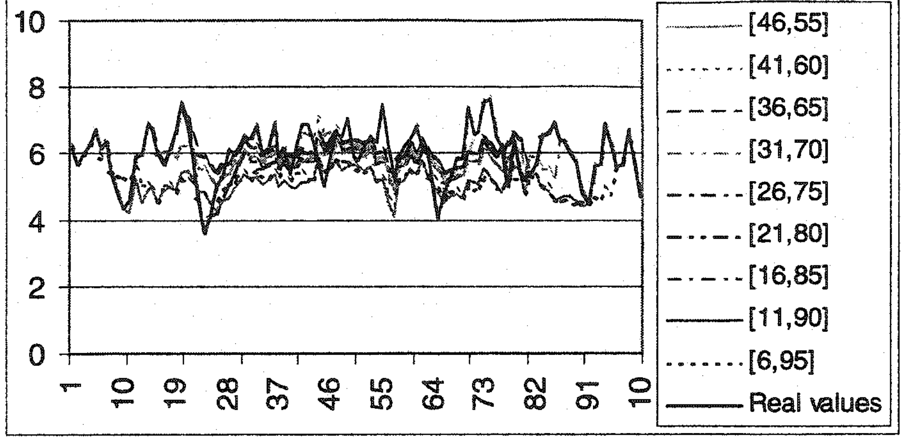


Figure B.23: Predictions of Y2 from the estimation of a GVAR(1) model based on simulated data

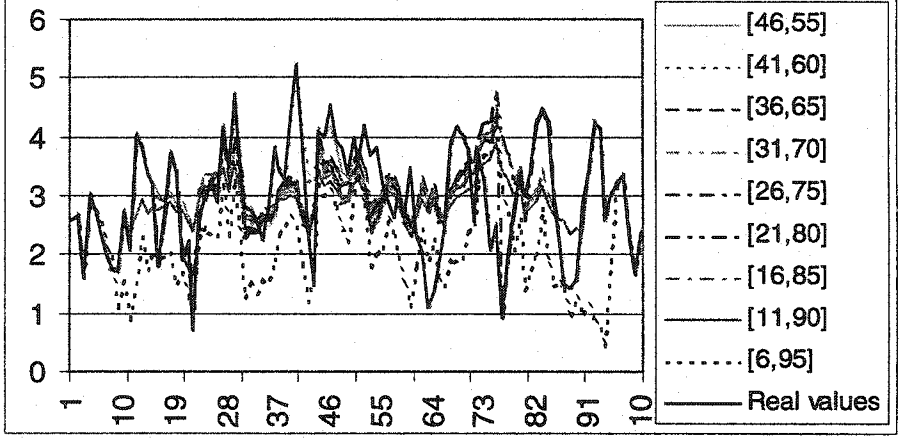


Figure B.24: Comparison of predictions of Y1 from the estimation of a GVAR(1) model based on simulated data on the [26,75] interval

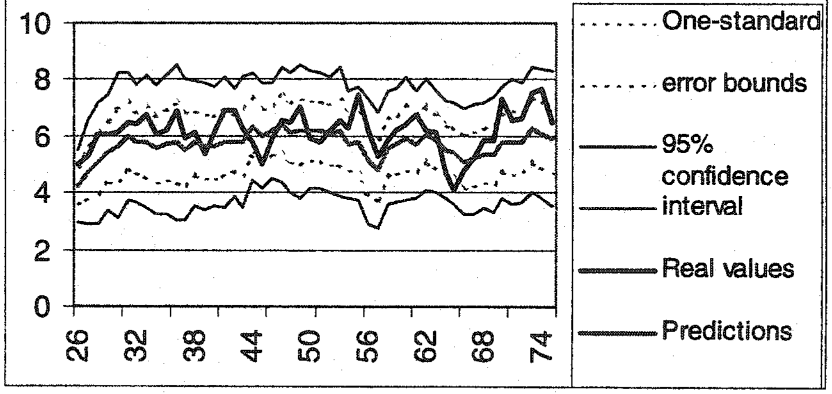


Figure B.25: Comparison of predictions of Y2 from the estimation of a GVAR(1) model based on simulated data on the [26,75] interval

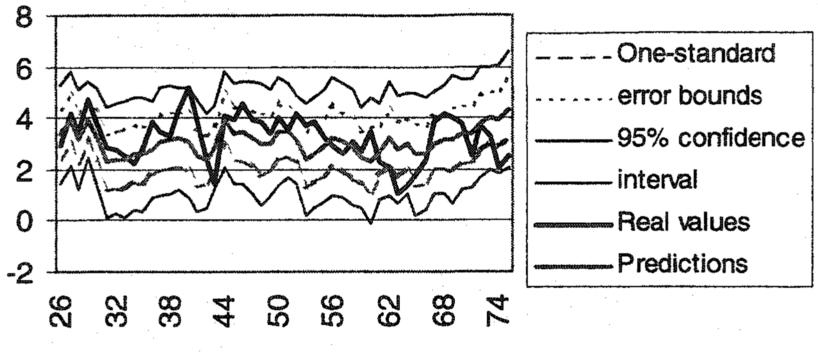


Figure B.26: Prediction errors on Y1 from the estimation of a GVAR(1) model based on simulated data

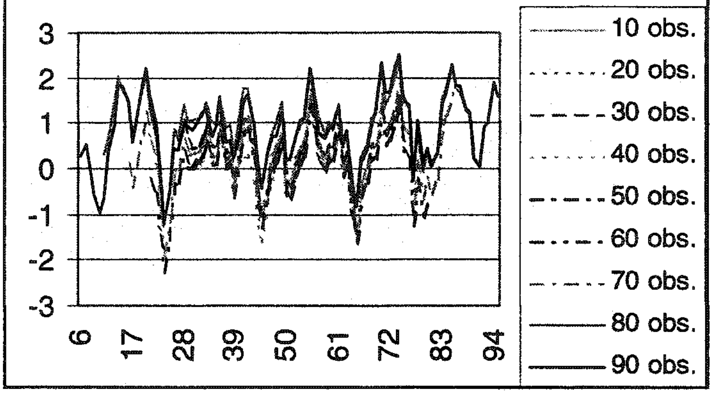
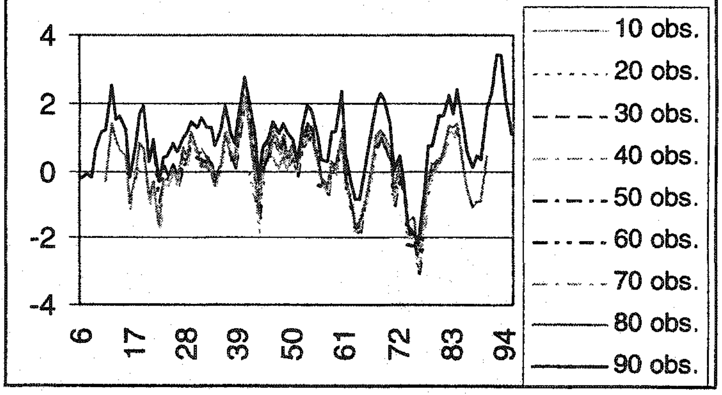


Figure B.27: Prediction errors on Y2 from the estimation of a GVAR(1) model based on simulated data



Appendix C

This appendix relates to the discussion of the empirical results proposed in Chapter 5. Firstly, we examine the sources and configuration of the independent variables appearing in the specifications described in Section 5.2 of Chapter 5. Secondly, we provide tables of estimation results for the selected model forms and specifications described in Section 5.3. Thirdly, series of tables and figures on prediction results are reported. These relate to the comparison of prediction results proposed in Section 5.4. By default, reported results correspond to those obtained by estimating the selected (GVAR(1)) model. However, some figures show comparisons with predictions obtained from alternative model types or specifications.

The registration data are drawn from the MVR.XLS file, which Natural Resources Canada (NRCan) kindly provided to us. However, the file had to be updated so as to include annual observations from 1995 on. The values produced by Statistics Canada from provincial registration files, as listed in its catalogues no. 53-219, are adjusted in the same way as in an earlier AMDC report (Boucher, 1998a) so as to exclude certain vehicle categories included in the registration registers of certain provinces.[1] Specifically, buses were separated from passenger automobiles where they were included in that class, and in certain cases other categories of vehicles were added to the passenger automobile class.

According to our most recent observations, this class corresponds to light vehicles, including cars and light trucks and vans, that are used for private purposes. As a result, passenger automobile registrations (REGIST) can be used to model the variables related to cars as well as light trucks and vans because they cannot be disaggregated by these types of vehicles.

[1] Note that the same types of adjustments were made for the additional 1995 and 1996 observations taken from Statistics Canada's catalogues.

The population is estimated on the basis of LFS performed on a monthly basis. It is taken from the CANSIM disk databank, which gives annual average estimates of the population aged 16 years or more, i.e. of legal driving age. The number of passenger automobile registrations divided by these estimates gives an estimate of per capita registrations (PCREGIST).

The gross domestic product of the public transit sector (GDPPUBT), based on the factors of production costs, was also drawn from the CANSIM disk databank, where it took the form of quarterly reports expressed in constant 1986 dollars. As in the case of the other variables extracted from the CANSIM disk databank, the series comprised raw data that had not been seasonally adjusted. For our study, the purpose of which is precisely to capture seasonal variations along with trends in the series requiring prediction, seasonally adjusted data would have been of no interest.

The price of new vehicles (VPRICE) is a weighted average of prices for the main exporting countries. The weights correspond to each exporting country's relative market share. The prices and relative shares for the exporting countries are drawn from the CANSIM disk databank. The series had to be deflated, because new vehicle prices are expressed in current dollars.

This was done using the monthly consumer price index (CPI) in 1986 dollars. The CPI series also comes from the CANSIM disk databank. The inflation adjustment process is always the same: a quantity expressed in current dollars is divided by the CPI, and the result is multiplied by 100. Since CPIs and new vehicle prices are surveyed on a monthly basis, the figures had to be aggregated to obtain quarterly averages.

Personal disposable income per capita (INCPC) is calculated using the annual gross domestic product, tax rates, and an estimate of the Canadian population. It is expressed in 1986 constant dollars and presented as an annual figure. The series is extracted from the CANSIM matrix of major indicators of the Canadian economy.

The quarterly figures of national income (QNATINC) and unleaded fuel price (QUNLFP) also come from CANSIM. Both series are adjusted to account for the inflation by using a quarterly average of the monthly CPIs in 1986 dollars.

The series of Canadian average temperatures (TEMP), expressed in degree-days, was provided by Environment Canada. Quarterly averages were calculated for the series. As the series begins only in 1981, it was only applied to the light truck and van specifications.

Canadian interest rates on personal consumer loans (IRATES) are extracted from the CANSIM databank. The chartered banks' interest rates on prime business loans come out of the Bank of Canada's web site which might be found at the address www.bankofcanada.ca. The U.S. interest rates series was found on the Internet at the address TedBos@UAB.edu, in the section "Economic Time Series, US Government, Federal Reserve, Board of Governors, Interest Rates." The series in question is entitled "Bank prime loan rate". All these series represent monthly averages and are expressed in nominal terms. They are aggregated on a quarterly basis using averages.

The on-road fuel consumption rates (FCR) come from the Desrosiers Automotives Consultants and were provided to us by NRCan. They are the annual series currently being used by NRCan in its energy demand prediction model (TEDM).

Net gasoline retail sales (FSALES) come from the CANSIM disk databank too. It is a monthly series expressed in cubic metres. First, the series is converted into litres using the rule 1 m^3 = 1000 litres. Second, the quarterly sum is calculated to bring it into line with the dependent variables.

Table C.1: Parameter estimates of the generalized vector autoregressive model with 1 lag of each variable in each equation relating to cars

Variable	**Vehicle stock (S)**			**Average distance ($\bar{D}$)**			**Fuel efficiency (E)**		
Statistics	Estim.	Std. err.	t-stat.	Estim.	Std. err.	t-stat.	Estim.	Std. err.	t-stat.
Constant	-0.825	1.386	-0.595	0.177	1.567	0.113	-1.725	2.226	-0.775
REGIST	0.462	0.159	2.911	----	----	----	----	----	----
PCREGIST	----	----	----	0.603	0.265	2.276	----	----	----
FCR	----	----	----	----	----	----	0.557	0.161	3.466
VPRICE	-0.007	0.006	-1.211	----	----	----	----	----	----
FALL	0.100	0.127	0.785	-1.197	0.120	-9.985	1.645	0.255	6.460
WIN	0.104	0.119	0.873	-1.199	0.097	-12.307	1.701	0.316	5.388
SPR	0.208	0.156	1.334	-0.452	0.131	-3.447	-0.471	0.322	-1.464
S_{-1}	0.478	0.147	3.243	0.012	0.038	0.320	0.063	0.138	0.456
$\angle_{-1}$	-0.028	0.085	-0.330	0.290	0.131	2.206	0.037	0.185	0.201
E_{-1}	-0.002	0.036	-0.060	-0.014	0.017	-0.859	0.472	0.126	3.729

Table C.2: Parameter estimates of the generalized vector autoregressive model with 1 lag of each variable in each equation relating to light trucks and vans

Variable	**Vehicle stock (S)**			**Average distance ($\bar{D}$)**			**Fuel efficiency (E)**		
Statistics	Estim.	Std. err.	t-stat.	Estim.	Std. err.	t-stat.	Estim.	Std. err.	t-stat.
Constant	-3.512	0.954	-3.681	-0.866	2.909	-0.298	-0.269	2.692	-0.100
REGIST	0.396	0.109	3.632	----	----	----	----	----	----
PCREGIST	----	----	----	0.775	0.556	1.394	----	----	----
FCR	----	----	----	----	----	----	0.731	0.230	3.180
VPRICE	-0.010	0.005	-2.023	----	----	----	----	----	----
FALL	0.158	0.087	1.809	-1.704	0.229	-7.435	1.803	0.416	4.339
WIN	-0.065	0.076	-0.852	-1.598	0.115	-13.899	1.279	0.461	2.772
SPR	0.032	0.082	0.389	-0.879	0.167	-5.252	-0.955	0.573	-1.665
S_{-1}	0.559	0.133	4.197	0.025	0.053	0.473	0.074	0.293	0.253
$\angle_{-1}$	0.002	0.044	0.056	0.339	0.185	1.836	-0.050	0.238	-0.211
E_{-1}	0.003	0.012	0.232	0.008	0.017	0.484	0.450	0.145	3.107

Table C.3: Parameter estimates of the error variance-covariance matrix for generalized vector autoregressive models or order 1

Variable	Cars			Light trucks and vans		
Statistics	Parameter estimates	Standard error	Student t-statistics	Parameter estimates	Standard error	Student t-statistics
Σ_{11}	0.065	0.017	3.846	0.023	0.011	2.002
Σ_{12}	-0.009	0.009	-0.997	-0.006	0.011	-0.523
Σ_{13}	0.041	0.026	1.563	-3x10^{-4}	0.019	-0.017
Σ_{22}	0.035	0.009	3.716	0.057	0.027	2.138
Σ_{23}	-0.038	0.022	-1.741	-0.022	0.041	-0.535
Σ_{33}	0.298	0.068	4.365	0.588	0.172	3.416
CPU time	8:58:06			3:02:31		

Table C.4 : Predictions for variables relating to private-use cars

Year-qtr	S: Average number of vehicles (units)	$\overline{D}$: Average distance travelled (km)	E: Weighted fuel consumption rates (litres/100 km)
79t4	7 365 300	4 050	17.2
80t1	7 410 900	3 320	18.6
80t2	7 504 000	4 070	15.5
80t3	7 408 300	4 560	15.0
80t4	7 384 300	3 570	17.7
81t1	7 135 200	3 290	17.3
81t2	7 529 800	4 160	15.0
81t3	7 326 900	4 330	15.4
81t4	7 626 800	3 440	16.3
82t1	7 537 700	3 170	17.6
82t2	7 228 800	4 160	14.1
82t3	7 406 200	4 660	13.5
82t4	7 570 000	3 800	15.2
83t1	7 510 800	3 610	15.4
83t2	7 337 200	4 130	13.7
83t3	7 281 900	4 690	13.2
83t4	7 203 700	3 730	14.4
84t1	7 432 700	3 440	15.3
84t2	7 299 300	4 080	13.3
84t3	7 612 100	4 710	12.9
84t4	7 692 600	3 820	14.0
85t1	7 684 700	3 470	14.9
85t2	8 431 000	4 150	13.5
85t3	7 821 600	4 580	12.0
85t4	7 880 000	3 810	13.8
86t1	8 176 600	3 310	14.3

Year-qtr	S: Average number of vehicles (units)	$\overline{D}$: Average distance travelled (km)	E: Weighted fuel consumption rates (litres/100 km)
86t2	8 128 100	4 170	12.1
86t3	7 988 200	4 790	12.3
86t4	7 828 900	4 050	13.2
87t1	8 386 300	3 750	13.1
87t2	8 542 000	4 310	11.3
87t3	8 212 800	5 090	11.4
87t4	7 996 800	3 850	12.7
88t1	8 578 300	3 930	13.0
88t2	8 701 900	4 130	11.7
88t3	8 384 600	5 190	10.9
88t4	8 703 200	3 920	12.5
89t1	8 789 200 [8 251 600,9 326 700]	3 720 [3 360,4 070]	12.9 [11.9,13.9]
89t2	8 988 400 [8 404 500,9 572 200]	4 400 [3 970,4 830]	10.9 [9.6,12.2]
89t3	8 816 400 [8 212 500,9 420 300]	5 080 [4 640,5 510]	10.5 [9.3,11.6]
89t4	8 836 700 [8 128 400,9 544 900]	4 080 [3 630,4 530]	11.9 [10.8,13.1]
90t1	8 994 400 [8 254 400,9 734 400]	3 790 [3 320,4 270]	12.5 [11.4,13.6]
90t2	9 194 200 [8 579 700,9 808 800]	4 450 [4 010,4 890]	10.6 [9.4,11.8]
90t3	9 044 200 [8 508 600,9 579 800]	5 120 [4 700,5 540]	10.2 [9.0,11.4]
90t4	9 031 600 [8 453 900,9 609 300]	4 130 [3 680,4 580]	11.6 [10.4,12.9]
91t1	9 081 100 [8 338 200,9 824 100]	3 730 [3 350,4 100]	12.2 [11.0,13.5]
91t2	9 232 900 [8 484 600,9 981 100]	4 350 [3 890,4 810]	10.3 [8.8,11.8]
91t3	9 078 300 [8 284 500,9 872 200]	5 010 [4 620,5 390]	9.9 [8.7,11.2]
91t4	9 032 900 [8 352 600,9 713 200]	4 010 [3 560,4 450]	11.4 [10.0,12.9]
92t1	9 170 900 [8 520 100,9 821 600]	3 690 [3 200,4 180]	12.0 [10.5,13.5]
92t2	9 340 100 [8 596 200,10 084 000]	4 330 [3 960,4 710]	10.1 [8.5,11.7]
92t3	9 202 800 [8 558 800,9 846 700]	5 000 [4 560,5 440]	9.7 [8.1,11.3]
92t4	9 208 600 [8 590 100,9 827 000]	4 000 [3 590,4 420]	11.2 [9.8,12.5]
93t1	9 334 500 [8 751 000,9 918 100]	3 690 [3 290,4 080]	11.8 [10.6,13.1]

Year-qtr	S: Average number of vehicles (units)	$\overline{D}$: Average distance travelled (km)	E: Weighted fuel consumption rates (litres/100 km)
93t2	9 525 600 [8 745 700,10 305 500]	4 330 [3 950,4 720]	9.9 [8.6,11.3]
93t3	9 380 900 [8 649 200,10 112 700]	5 000 [4 630, 5 380]	9.6 [8.4,10.7]
93t4	9 413 500 [8 682 300,10 144 600]	4 000 [3 550,4 460]	11.0 [9.9,12.2]
94t1	9 583 000 [8 761 600,10 404 400]	3 690 [3 270,4 120]	11.7 [10.4,13.0]
94t2	9 765 300 [9 115 600,10 415 100]	4 340 [3 780,4 900]	9.8 [8.5,11.2]
94t3	9 590 300 [8 983 900,10 196 600]	5 010 [4 510,5 500]	9.5 [8.2,10.7]
94t4	9 483 900	3 970	10.6
95t1	9 220 300	3 760	10.8
95t2	10 097 600	4 100	10.1
95t3	9 772 000	5 410	9.7
95t4	9 520 100	4 050	11.3
96t1	9 892 800	3 460	11.2
96t2	10 130 900	4 100	10.0
96t3	9 962 900	4 920	9.2

Legend: Numbers in the shaded area correspond to predictions, while other entries refer to adjusted survey-based estimates. In the shaded area, the second entry represents a 95% confidence interval on predictions. Numbers are rounded to meet Statistics Canada's precision criteria. Precisely, estimates of the number of vehicles are rounded to the nearest hundred, the average distance estimates are rounded to the nearest ten and estimates of the average fuel consumption rate are rounded to the first decimal point.

Table C.5: Predictions for variables relating to private-use light trucks and vans

Year-qtr	S: Average number of vehicles (units)	$\overline{D}$: Average distance travelled (km)	E: Weighted fuel consumption rates (litres/100 km)
81t4	1 448 100	3 810	21.2
82t1	1 392 700	3 380	23.4
82t2	1 407 600	4 340	19.6
82t3	1 435 900	5 020	18.8
82t4	1 599 000	3 920	20.7
83t1	1 560 700	3 660	21.1
83t2	1 577 500	4 060	18.5
83t3	1 422 700	5 150	18.5
83t4	1 530 100	3 580	20.2
84t1	1 452 400	3 530	20.6
84t2	1 460 100	4 080	17.9
84t3	1 454 700	4 840	17.5
84t4	1 581 100	4 100	19.3
85t1	1 691 000	3 560	19.6
85t2	1 756 100	4 310	16.7
85t3	1 689 100	5 020	16.7
85t4	1 739 600	3 720	18.9
86t1	1 839 800	3 620	18.4
86t2	1 879 700	4 230	16.8
86t3	1 899 700	5 480	16.2
86t4	1 845 400	3 880	19.1
87t1	1 884 100	3 730	17.1
87t2	1 909 300	4 170	15.8
87t3	2 056 000	5 710	16.0
87t4	2 072 600	4 170	17.4
88t1	2 084 600 [1 746 400,2 422 800]	3 830 [3 330,4 330]	17.3 [15.7,19.0]
88t2	2 236 200 [1 840 300,2 632 000]	4 440 [3 800,5 080]	15.1 [13.3,17.0]
88t3	2 226 600 [1 852 400,2 600 700]	5 500 [4 730,6 270]	15.0 [13.0,17.1]
88t4	2 430 000 [2 010 700,2 849 500]	4 150 [3 340,4 970]	16.8 [15.0,18.6]
89t1	2 407 200 [2 004 000,2 810 400]	3 860 [3 230,4 500]	16.9 [15.6,18.1]

Year-qtr	S: Average number of vehicles (units)	$\overline{D}$: Average distance travelled (km)	E: Weighted fuel consumption rates (litres/100 km)
89t2	2 565 200 [2 052 400,3 077 900]	4 500 [3 690,5 300]	14.6 [12.6,16.6]
89t3	2 549 400 [2 075 800,3 023 000]	5 570 [4 760,6 380]	14.6 [12.9,16.3]
89t4	2 726 600 [2 202 700,3 250 600]	4 230 [3 330,5 120]	16.3 [14.7,18.0]
90t1	2 710 300 [2 253 300,3 167 200]	3 920 [3 060,4 770]	16.5 [14.7,18.3]
90t2	2 810 600 [2 346 500,3 274 700]	4 550 [4 000,5 110]	14.3 [12.6,16.0]
90t3	2 814 900 [2 423 600,3 206 100]	5 600 [5 000,6 200]	14.2 [12.5,15.9]
90t4	2 933 100 [2 560 300,3 305 900]	4 270 [3 500,5 040]	15.9 [14.1,17.8]
91t1	2 845 400 [2 437 300,3 253 600]	3 830 [3 080,4 580]	16.1 [14.4,17.8]
91t2	2 915 800 [2 516 300,3 315 200]	4 400 [3 680,5 120]	14.0 [11.8,16.1]
91t3	2 909 100 [2 478 500,3 339 600]	5 460 [4 750,6 160]	13.9 [11.9,16.0]
91t4	2 981 700 [2 523 000,3 440 400]	4 100 [3 540,4 660]	15.7 [13.7.17.6]
92t1	2 936 400 [2 491 000,3 381 800]	3 760 [3 040,4 480]	15.8 [14.0,17.7]
92t2	3 005 900 [2 559 500,3 452 200]	4 360 [3 710,5 010]	13.7 [11.8,15.6]
92t3	3 001 700 [2 488 200,3 515 100]	5 430 [4 740,6 120]	13.7 [11.7,15.7]
92t4	3 156 700 [2 718 400,3 594 900]	4 090 [3 390,4 790]	15.4 [13.8,17.1]
93t1	3 095 600 [2 650 900,3 540 200]	3 760 [3 130,4 400]	15.6 [13.9,17.4]
93t2	3 193 100 [2 686 400,3 699 900]	4 360 [3 700,5 010]	13.5 [11.5,15.5]
93t3	3 188 600 [2 703 600,3 673 600]	5 440 [4 820,6 050]	13.5 [11.8,15.2]
93t4	3 379 100 [2 878 700,3 879 400]	4 090 [3 540,4 640]	15.3 [13.7,16.8]
94t1	3 385 100 [2 990 600,3 779 500]	3 760 [3 190,4 320]	15.5 [13.6,17.4]
94t2	3 477 400 [3 023 800,3 931 100]	4 360 [3 690,5 040]	13.4 [11.5,15.3]
94t3	3 432 500 [3 030 100,3 834 900]	5 430 [4 800,6 070]	13.3 [11.1,15.6]
94t4	3 845 400	4 030	14.8

Year-qtr	S: Average number of vehicles (units)	$\overline{D}$: Average distance travelled (km)	E: Weighted fuel consumption rates (litres/100 km)
95t1	3 376 500	3 920	14.0
95t2	3 723 800	4 160	12.9
95t3	3 329 000	5 730	14.1
95t4	3 612 600	4 200	14.7
96t1	3 618 000	3 320	15.2
96t2	3 668 500	4 240	13.6
96t3	3 836 300	5 230	13.5

Legend: Numbers in the shaded area correspond to predictions, while other entries refer to adjusted survey-based estimates. In the shaded area, the second entry represents a 95% confidence interval on predictions. Numbers are rounded to meet Statistics Canada's precision criteria. Precisely, estimates of the number of vehicles are rounded to the nearest hundred, the average distance estimates are rounded to the nearest ten and estimates of the average fuel consumption rate are rounded to the first decimal point.

Figure C.1: Comparison of predictions of the average number of cars obtained by estimating different model types

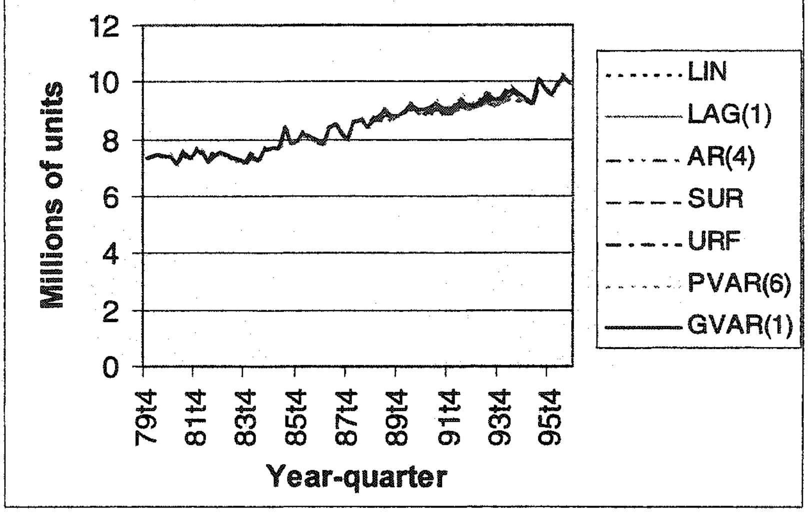


Figure C.2: Comparison of predictions of the average number of light trucks and vans obtained by estimating different model types

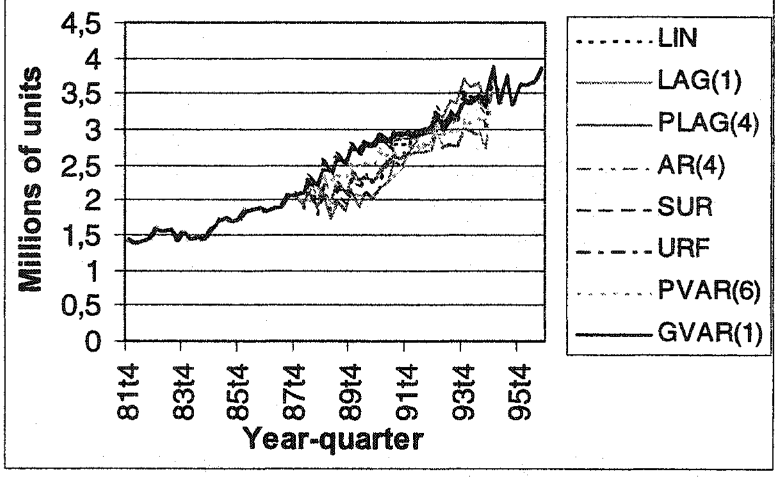


Figure C.3: Comparison of predictions of the average distance travelled by car obtained by estimating different model types

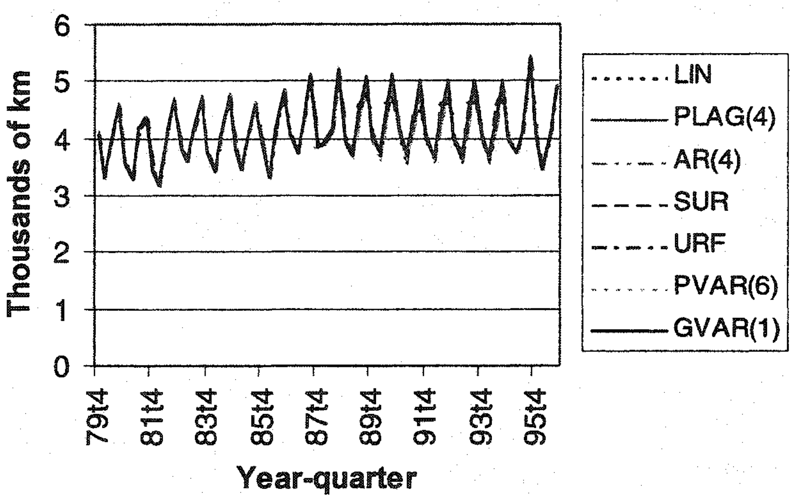


Figure C.4: Comparison of predictions of the average distance travelled by light trucks and vans obtained by estimating different model types

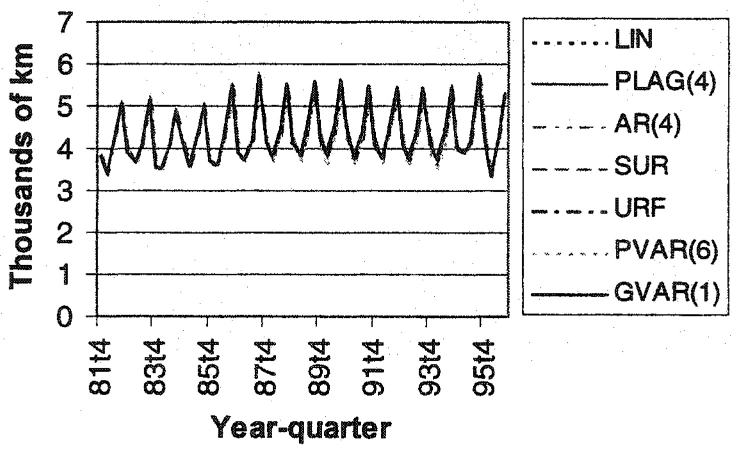


Figure C.5: Comparison of predictions of the average distance travelled by vehicle type

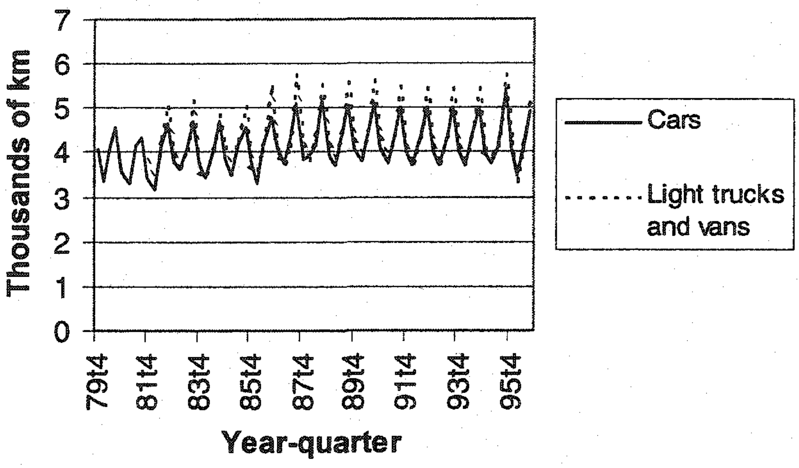


Figure C.6: Comparison of predictions of the average weighted fuel consumption rates of cars obtained by estimating different model types

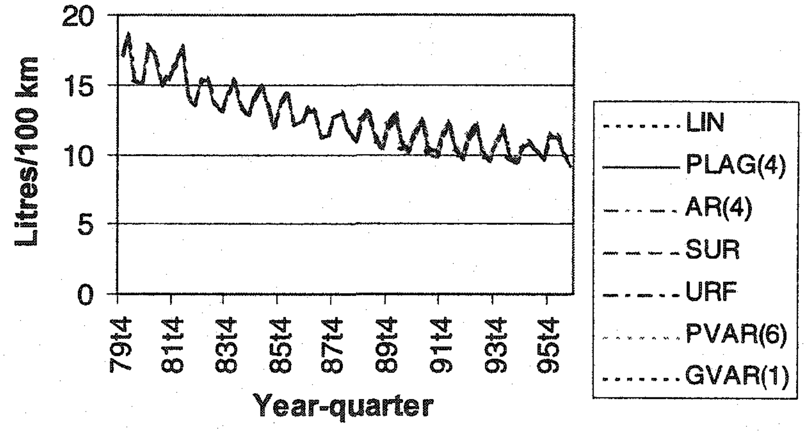


Figure C.7: Comparison of predictions of the average weighted fuel consumption rates of light trucks and vans obtained by estimating different model types

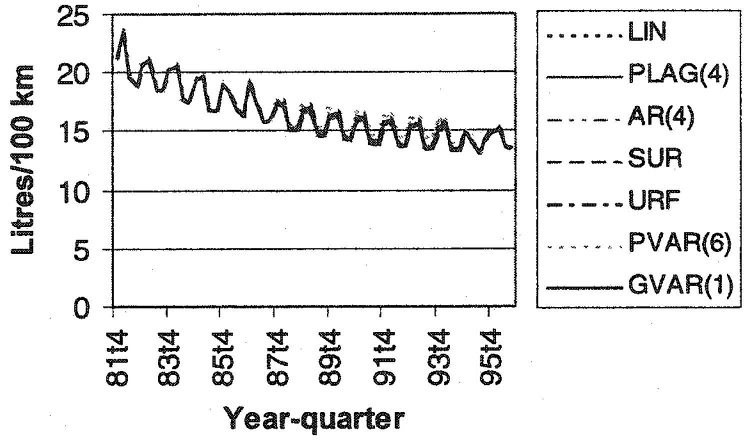
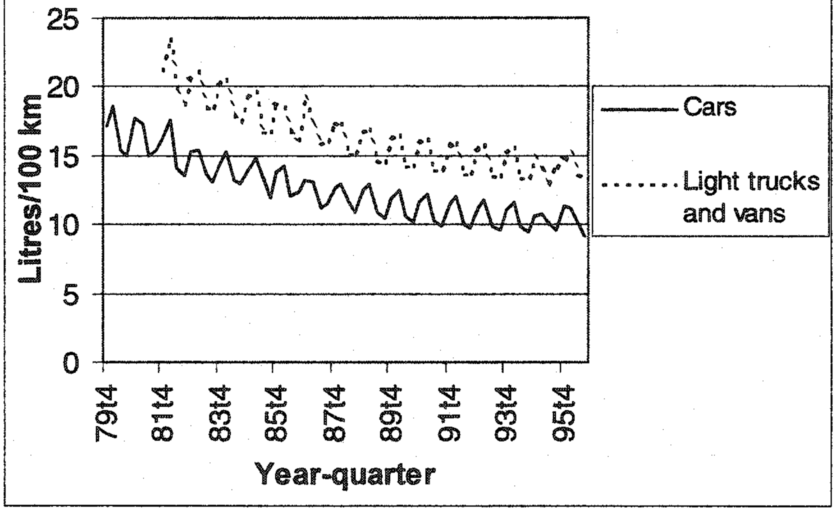


Figure C.8: Comparison of predictions of the weighted fuel consumption rates by vehicle type



Appendix D

This appendix relates to the stability and robustness checks of the model specification performed in Chapter 6. Firstly, the root mean squared error (RMSE) results obtained by enlarging the bounds of the prediction period to encompass parts of the survey samples are presented in Tables D.1-D.6. Secondly, a selection of figures are provided to support the argument of Section 6.2 regarding the choice of the model type. Thirdly, Tables D.7-D.8 summarize the information criteria (AIC and BIC) used, in Section 6.3, to determine the number of lagged dependent variables included in the GVAR model formulations for cars and light trucks/vans, respectively. Competing model specifications considered, in Section 6.4, for predicting the average number of light trucks and vans are described in Table D.9. Resulting predictions are compared in Figure D.6. Tables D.10-D.12 provide the RMSE results used in selecting, among them, the best formulation for prediction purposes. The last figure illustrates the results of the sensitivity analysis to the prior choice performed in Section 6.6.

Table D.1 **Root mean squared errors on predictions from the different model types obtained by removing observations at the end of the FCS sampling period**

Model	1 obs.	2 obs.	3 obs.	4 obs.	5 obs.	6 obs.	7 obs.	8 obs.
LIN	0.477	0.414	0.286	0.200	0.164	0.127	0.118	0.117
LAG(1)	0.348	0.254	0.164	0.129	0.097	0.115	0.158	0.159
AR(4)	0.473	0.630	0.336	0.311	0.543	0.388	0.382	0.486
SUR	0.100	0.109	0.076	0.076	0.060	0.053	0.063	0.092
PVAR(1)	0.212	0.099	0.065	0.063	0.132	0.121	0.094	0.056
PVAR(2)	0.178	0.094	0.075	0.046	0.132	0.074	0.092	0.081
PVAR(3)	0.195	0.095	0.061	0.059	0.128	0.133	0.113	6.042
PVAR(4)	0.168	0.124	0.066	0.051	0.138	0.091	---	---
PVAR(5)	0.241	---	---	0.053	---	---	---	---
GVAR(1)	0.057	0.092	0.056	0.036	0.056	0.066	0.074	0.030
GVAR(2)	0.039	0.088	0.052	0.029	0.077	0.072	0.089	0.092
GVAR(3)	0.026	0.077	0.045	0.031	0.062	0.052	0.061	0.104
GVAR(4)	0.053	0.077	0.054	0.034	0.033	0.099	0.063	---
GVAR(5)	0.026	0.092	0.046	---	---	---	---	---

Legend: The second line corresponds to the simple linear regression model and the third adds a four-period lagged dependent variable to the former. The fourth line corresponds to the simple linear regression model with AR(4) errors. The fifth line provides results for the SUR model. PVAR and GVAR refer, respectively, to the pure and generalized VAR models. The numbers in parentheses give the numbers of lags of the dependent variables included in the model specification. Dashes stand for results that could not be computed either because the estimation process did not converge or because the number of degrees was insufficient for estimation.

Table D.2 **Root mean squared errors on predictions from the different model types obtained by removing observations at the beginning of the NaPVUS sampling period**

Model	**1 obs.**	**2 obs.**	**3 obs.**	**4 obs.**	**5 obs.**	**6 obs.**	**7 obs.**	**8 obs.**
LIN	1.821	1.083	0.948	0.787	0.725	0.658	0.619	0.618
LAG(1)	1.844	1.106	0.938	0.783	0.721	0.654	0.614	0.587
AR(4)	1.646	0.976	0.822	0.698	0.624	0.598	0.578	0.604
SUR	1.565	0.932	0.783	0.620	0.538	0.477	0.443	0.607
PVAR(1)	1.702	1.069	0.900	0.738	0.600	0.612	0.529	0.539
PVAR(2)	1.696	---	0.868	0.718	0.695	5.001	0.596	0.540
PVAR(3)	1.717	---	---	0.751	0.606	0.542	0.582	0.507
PVAR(4)	1.675	1.040	0.437	0.774	0.711	0.578	0.613	0.521
PVAR(5)	1.643	---	---	---	---	---	---	0.486
GVAR(1)	1.657	0.993	0.824	0.663	0.569	0.486	0.439	0.503
GVAR(2)	1.657	0.995	---	0.655	0.557	0.507	0.445	0.513
GVAR(3)	1.644	0.977	0.805	0.636	0.541	0.478	0.670	0.482
GVAR(4)	1.632	0.978	0.807	0.689	0.526	0.452	0.409	0.512
GVAR(5)	1.641	0.972	0.790	0.626	0.532	25.934	0.379	---

Legend: The second line corresponds to the simple linear regression model and the third adds a four-period lagged dependent variable to the former. The fourth line corresponds to the simple linear regression model with AR(4) errors. The fifth line provides results for the SUR model. PVAR and GVAR refer, respectively, to the pure and generalized VAR models. The numbers in parentheses give the numbers of lags of the dependent variables included in the model specification. Dashes stand for results that could not be computed either because the estimation process did not converge or because the number of degrees was insufficient for estimation.

Table D.3: **Root mean squared errors on predictions from the different model types obtained by adding observations at both ends of the prediction period**

Model	**1 obs.**	**2 obs.**	**3 obs.**	**4 obs.**	**5 obs.**	**6 obs.**	**7 obs.**	**8 obs.**
LIN	0.990	0.633	0.542	0.452	0.435	0.371	0.320	0.327
LAG(1)	0.965	0.606	0.516	0.426	0.387	0.330	0.274	0.309
AR(4)	0.639	0.376	0.329	0.309	0.286	0.279	0.238	0.225
SUR	0.797	0.499	0.413	0.321	0.294	0.261	0.237	0.317
PVAR(1)	0.818	0.567	0.418	0.405	0.268	0.484	0.302	0.316
PVAR(2)	0.907	0.546	13.860	0.360	0.236	0.286	0.294	0.345
PVAR(3)	0.787	0.969	0.416	0.360	---	0.339	---	0.307
PVAR(4)	0.842	0.419	0.419	---	0.330	0.315	0.201	---
PVAR(5)	---	0.567	---	---	---	---	---	---
GVAR(1)	0.813	0.535	0.434	0.337	0.294	0.244	0.774	0.279
GVAR(2)	0.813	0.526	0.459	0.359	0.420	0.511	0.775	0.231
GVAR(3)	0.785	0.514	0.441	0.320	0.223	0.556	---	---
GVAR(4)	0.782	0.502	0.421	---	0.156	---	---	---
GVAR(5)	0.791	0.517	---	---	---	---	---	---

Legend: The second line corresponds to the simple linear regression model and the third adds a four-period lagged dependent variable to the former. The fourth line corresponds to the simple linear regression model with AR(4) errors. The fifth line provides results for the SUR model. PVAR and GVAR refer, respectively, to the pure and generalized VAR models. The numbers in parentheses give the numbers of lags of the dependent variables included in the model specification. Dashes stand for results that could not be computed either because the estimation process did not converge or because the number of degrees was insufficient for estimation.

Table D.4 Models by ascending order of root mean squared errors on predictions obtained by removing observations at the end of the FCS sampling period

1 obs.	2 obs.	3 obs.	4 obs.	5 obs.	6 obs.	7 obs.	8 obs.
GVAR(5)	GVAR(4)	GVAR(3)	GVAR(2)	GVAR(4)	GVAR(3)	GVAR(3)	GVAR(1)
GVAR(3)	GVAR(3)	GVAR(5)	GVAR(3)	GVAR(1)	GVAR(4)	SUR	PVAR(1)
GVAR(2)	GVAR(2)	GVAR(2)	GVAR(4)	SUR	GVAR(1)	GVAR(4)	PVAR(3)
GVAR(4)	GVAR(5)	GVAR(4)	GVAR(1)	GVAR(3)	GVAR(2)	GVAR(1)	PVAR(2)
GVAR(1)	GVAR(1)	GVAR(1)	PVAR(2)	GVAR(2)	SUR	GVAR(2)	GVAR((2)
SUR	PVAR(2)	PVAR(3)	PVAR(4)	LAG(1)	PVAR(2)	PVAR(2)	SUR
PVAR(4)	PVAR(3)	PVAR(1)	PVAR(5)	PVAR(3)	PVAR(4)	PVAR(1)	GVAR(3)
PVAR(2)	PVAR(1)	PVAR(4)	PVAR(3)	PVAR(1)	LAG(1)	PVAR(3)	LIN
PVAR(3)	PVAR(4)	PVAR(2)	PVAR(1)	PVAR(2)	PVAR(1)	LIN	LAG(1)
PVAR(1)	SUR	SUR	SUR	PVAR(4)	LIN	LAG(1)	AR(4)
PVAR(5)	LAG(1)	LAG(1)	LAG(1)	LIN	PVAR(3)	AR(4)	
LAG(1)	LIN	LIN	LIN	AR(4)	AR(4)		
AR(4)	AR(4)	AR(4)	AR(4)				
LIN							

Table D.5 Models by ascending order of root mean squared errors on predictions obtained by removing observations at the beginning of the NaPVUS sampling period

1 obs.	2 obs.	3 obs.	4 obs.	5 obs.	6 obs.	7 obs.	8 obs.
SUR	SUR	PVAR(4)	SUR	GVAR(4)	GVAR(4)	GVAR(5)	GVAR(3)
GVAR(4)	GVAR(5)	SUR	GVAR(5)	GVAR(5)	SUR	GVAR(4)	PVAR(5)
GVAR(5)	AR(4)	GVAR(5)	GVAR(3)	SUR	GVAR(3)	GVAR(1)	GVAR(1)
PVAR(5)	GVAR(3)	GVAR(3)	GVAR(2)	GVAR(3)	GVAR(1)	SUR	PVAR(3)
GVAR(3)	GVAR(4)	GVAR(4)	GVAR(1)	GVAR(2)	GVAR(2)	GVAR(2)	GVAR(4)
AR(4)	GVAR(1)	AR(4)	GVAR(4)	GVAR(1)	PVAR(3)	PVAR(1)	GVAR(2)
GVAR(1)	GVAR(2)	GVAR(1)	AR(4)	PVAR(1)	PVAR(4)	AR(4)	PVAR(4)
GVAR(2)	PVAR(4)	PVAR(2)	PVAR(2)	PVAR(3)	AR(4)	PVAR(3)	PVAR(1)
PVAR(4)	PVAR(1)	PVAR(1)	PVAR(1)	AR(4)	PVAR(1)	PVAR(2)	PVAR(2)
PVAR(2)	LIN	LAG(1)	PVAR(3)	PVAR(2)	LAG(1)	PVAR(4)	LAG(1)
PVAR(1)	LAG(1)	LIN	PVAR(4)	PVAR(4)	LIN	LAG(1)	AR(4)
PVAR(3)			LAG(1)	LAG(1)		LIN	SUR
LIN			LIN	LIN		GVAR(3)	LIN
LAG(1)							

Table D.6 **Models by ascending order of root mean squared errors on predictions obtained by adding observations at both ends of the prediction period**

1 obs.	2 obs.	3 obs.	4 obs.	5 obs.	6 obs.	7 obs.	8 obs.
AR(4)	AR(4)	AR(4)	AR(4)	GVAR(4)	GVAR(1)	PVAR(4)	AR(4)
GVAR(4)	SUR	SUR	GVAR(3)	GVAR(3)	SUR	SUR	GVAR(2)
GVAR(3)	GVAR(4)	PVAR(3)	SUR	PVAR(2)	AR(4)	AR(4)	GVAR(1)
PVAR(3)	GVAR(3)	PVAR(1)	GVAR(1)	PVAR(1)	PVAR(2)	LAG(1)	PVAR(3)
GVAR(5)	GVAR(5)	PVAR(4)	GVAR(2)	AR(4)	PVAR(4)	PVAR(2)	LAG(1)
SUR	GVAR(2)	GVAR(4)	PVAR(2)	GVAR(1)	LAG(1)	PVAR(1)	PVAR(1)
GVAR(1)	GVAR(1)	GVAR(1)	PVAR(3)	SUR	PVAR(3)	LIN	SUR
GVAR(2)	PVAR(2)	GVAR(3)	PVAR(1)	PVAR(4)	LIN	GVAR(1)	LIN
PVAR(1)	PVAR(5)	GVAR(2)	LAG(1)	LAG(1)	PVAR(1)	GVAR(2)	PVAR(2)
PVAR(4)	PVAR(1)	LAG(1)	LIN	GVAR(2)	GVAR(2)		
PVAR(2)	PVAR(4)	LIN		LIN	GVAR(3)		
LAG(1)	LAG(1)						
LIN	LIN						
	PVAR(3)						

Figure D.1: Predictions of the simple linear regression model with AR(4) errors obtained by removing observations from the NaPVUS

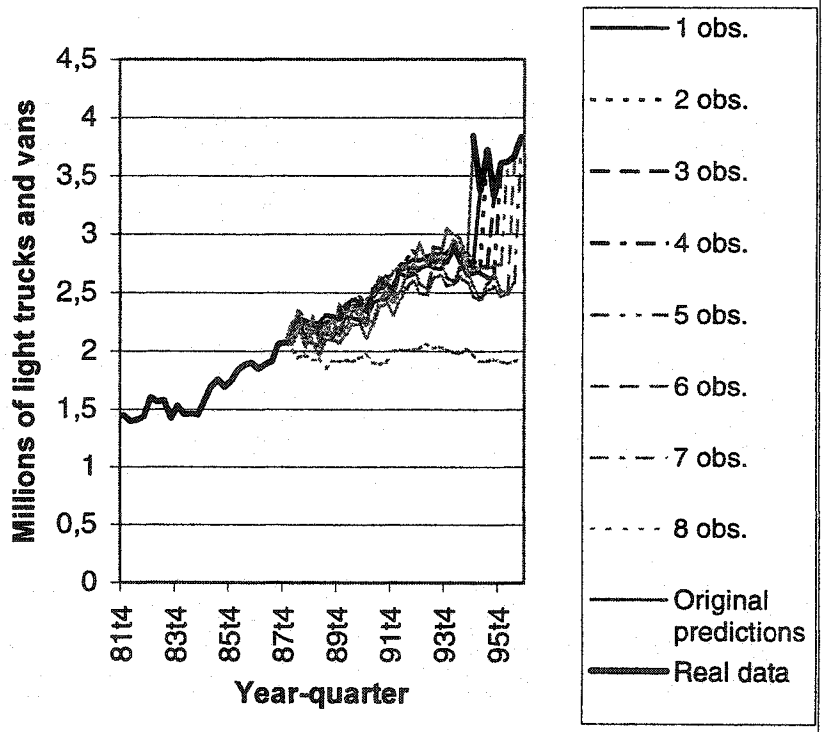


Figure D.2: Predictions of the simple linear regression model with AR(4) errors obtained by removing observations from the FCS

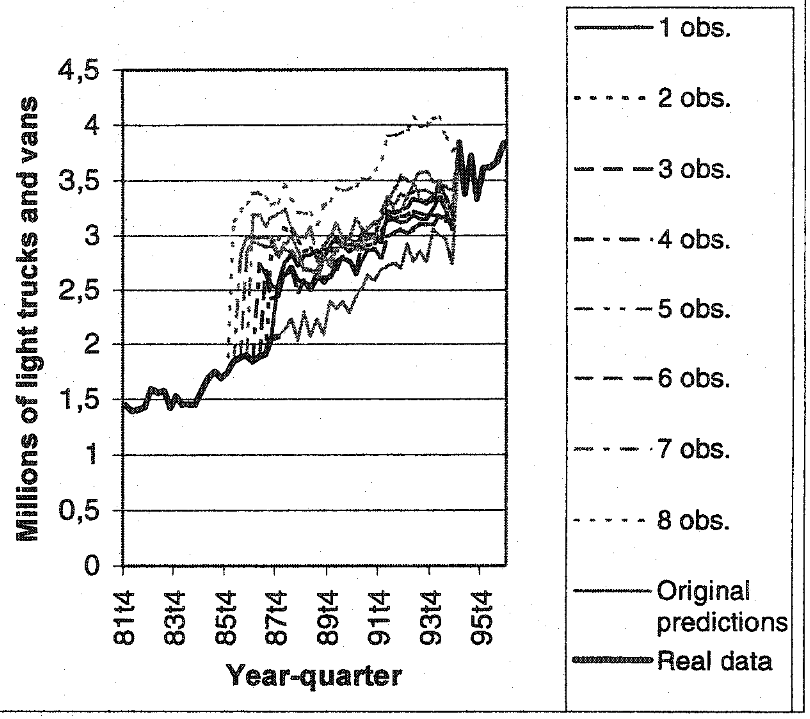


Figure D.3: Comparison of the predictions of the linear regression model with AR(4) errors obtained by removing one observation from the survey sample

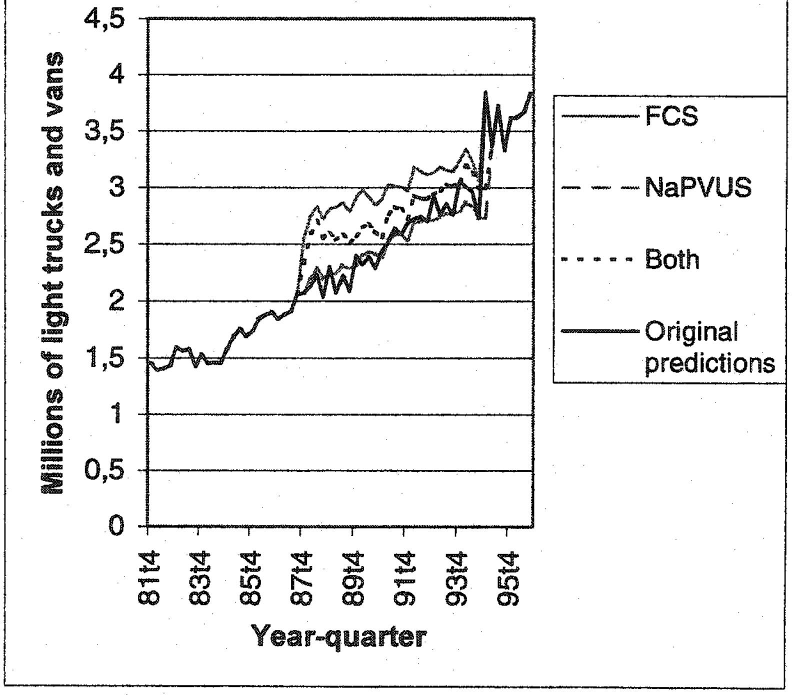


Figure D.4: Predictions of the GVAR(1) model obtained by removing observations from the NaPVUS

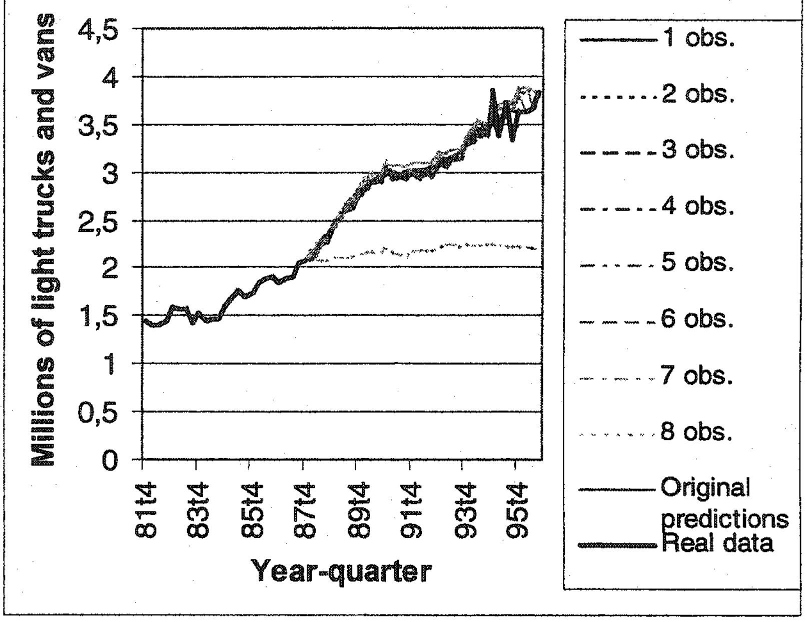


Figure D.5: Comparison of the competing models' performance at predicting two observations from the NaPVUS

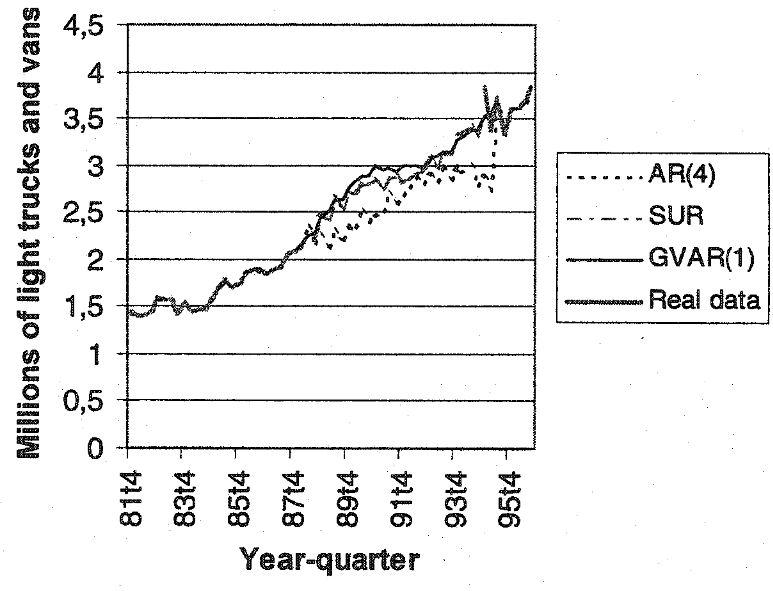


Table D.7 **Information criteria to determine the order (p) of generalized vector autoregressive models GVAR(p) for the cars variables**

p	Δ BIC	Δ AIC	LR	ρ_1	ρ_2	T-(p+1)
1	-37.735	-18.028	-0.028	25	34	66
2	-37.580	-18.011	-0.011	34	43	65
3	-37.171	-17.741	0.259	43	52	64
4	-37.559	-18.271	-0.271	52	61	63
5	-36.969	-17.824	0.176	61	70	62

Note: Because of degrees of freedom constraints, we were unable to estimate the model with 6 lags.

Table D.8 **Information criteria to determine the order (p) of generalized vector autoregressive models GVAR(p) for the light trucks and vans variables**

p	Δ BIC	Δ AIC	LR	ρ_1	ρ_2	T-(p+1)
1	-36.318	-17.874	0.126	25	34	58
2	-35.049	-16.661	1.339	34	43	57
3	-40.282	-22.053	-4.053	43	52	56
4	-36.048	-17.982	0.018	52	61	55
1/3	-54.930	-18.155	17.845	22	40	57

Note: Because of degrees of freedom constraints, we were unable to estimate the model with 6 lags.

Table D.9: **Exogenous explanatory variables involved in alternative specifications for variables relating to light trucks and vans**

Dep. var.	Vehicle stock (S)				Average distance ($\bar{D}$)				Fuel efficiency (E)			
Indep. var.	Var 1	Var 2	Var 3	Var 4	Var 1	Var 2	Var 3	Var 4	Var 1	Var 2	Var 3	Var 4
Constant	X	X	X	X	X	X	X	X	X	X	X	X
REGIST	X	X										
PCREGIST					X	X						
FCR									X	X		
VPRICE	X	X		X								
IRATES	X											
ARVDPC			X				X				X	
QNATINC				X				X				X
QCPI			X				X				X	
QUNLFP				X				X				X
QFSALES					X				X			
FALL	X	X	X	X	X	X	X	X		X	X	X
WIN	X	X	X	X	X	X	X	X		X	X	X
SPR	X	X	X	X	X	X	X	X		X	X	X
TEMP									X			

Figure D.6: Comparison of predictions of the average number of light trucks and vans based on different specifications of the simple linear regression model

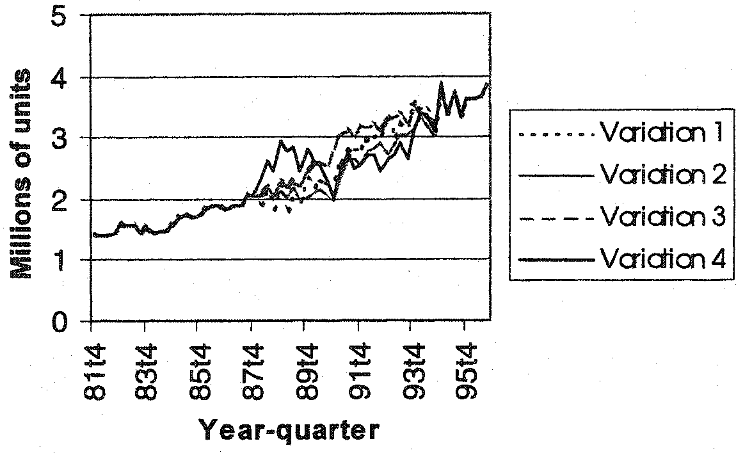


Table D.10: Root mean squared errors on predictions from the GVAR(1) model with various specifications obtained by removing observations at the end of the FCS sampling period

Nu. of missing obs.	Variation 1	Variation 2	Variation 3
1 observation	0.104	0.057	0.146
2 observations	0.128	0.092	0.070
3 observations	0.073	0.056	0.048
4 observations	0.072	0.036	0.036
5 observations	0.065	0.056	0.088
6 observations	0.041	0.066	0.059
7 observations	0.042	0.074	0.050
8 observations	0.037	0.030	0.039

Table D.11: Root mean squared errors on predictions from the GVAR(1) model with various specifications obtained by removing observations at the beginning of the NaPVUS sampling period

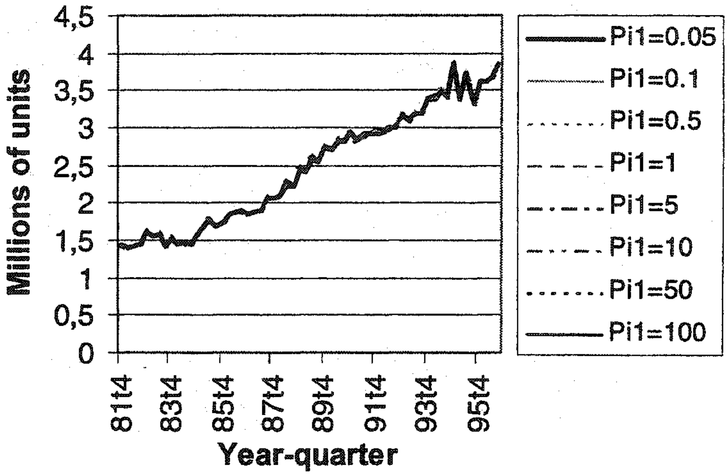
Nu. of missing obs.	Variation 1	Variation 2	Variation 3
1 observation	1.720	1.657	1.778
2 observations	1.019	0.993	1.061
3 observations	0.838	0.824	0.908
4 observations	0.666	0.663	0.720
5 observations	0.585	0.569	0.630
6 observations	0.503	0.486	0.593
7 observations	0.446	0.439	0.568
8 observations	0.541	0.503	0.462

Table D.12: Root mean squared errors on predictions from the GVAR(1) model with various specifications obtained by adding observations at both ends of the prediction period

Nu. of missing obs.	Variation 1	Variation 2	Variation 3
1 observation	0.823	0.813	0.842
2 observations	0.542	0.535	0.582
3 observations	0.458	0.434	0.434
4 observations	0.353	0.337	0.341
5 observations	0.246	0.294	0.329
6 observations	0.252	0.244	0.219
7 observations	----	0.774	----
8 observations	0.137	0.279	0.307

Note: RMSE results are not reported when the estimation process has not converged.

Figure D.7: Results of the sensitivity analysis to the prior choice on predictions of the number of light trucks and vans stock derived from the GVAR(1) model



Vita

Name: Nathalie Boucher

Place and Year of birth Lévis, Qc, 1966

Education:

Laval University, 1987-1990
B.A. (Mathematics) 1990

Laval University, 1990-1991
Preparatory stage to M.A. (Economics) 1991

Laval University, 1991-1993
M.A. (Economics) 1993

Queen's University, 1993-2002
Ph.D. 2002

Experience:

Analyst, Automobile Mobility Data Compendium (AMDC), Laval University, 1997-2000

Teachers' assistant, Department of Economics, Laval University, 1991-1993

Research assistant, Department of Economics, Queen's University, 1993-1994

Teachers' assistant, Department of Economics, Queen's University, 1994-1995

Research assistant, School of Policy Studies, Queen's University, 1996-1998, 2002

Analyst, AMDC, Laval University, 1997-2000

Conference at the Canadian Economics Association's (CEA) Annual Meeting, 1999

Advising expert, AMDC, Laval University, 2000-2002

Advising expert, Office of Energy Efficiency (OEE), Natural Resources Canada (NRCan), 2000-2002

Executive Director, AMDC, Laval University, since 2002

Awards:

University Scholarship, Group of Research on Economics of Energy and Natural Resources, Department of Economics, Laval University, 1991-1993

University Excellence Prize for the best M.A. essay (type B), Laval University, 1995

Bursary, Fonds pour la Formation des Chercheurs et l'Aide à la Recherche (FCAR), Québec Government, 1993-1997

Travel Grant, CEA, Annual Meeting, 1999

Publications:

Bonin, Sylvie and Boucher, Nathalie: Imputation des carnets incomplets de l'Enquête nationale sur l'utilisation des véhicules privés. AMCD. Final report. Document no. N97-03f. prepared for the OEE of NRCan as part of the 1997-1998 work plan. Laval University, Sainte Foy, 29 pages (1997)

Distinctions entre les nombres de véhicules à usage privé et commercial au Canada, 1980-1996. AMDC. Final report. Document prepared for the OEE of NRCan as part of the 1997-1998 work plan. Laval University, Sainte Foy. 65 pages (1998)

Étude des facteurs de marché influençant la consommation d'énergie dans le secteur du transport privé de passagers : 1990-1995. AMDC. Final report. Document no. N97-05f prepared for the OEE of NRCan as part of the 1997-1998 work plan. Laval University, Sainte Foy. 66 pages (1998)

Méthodologie proposée pour l'estimation de données nationales sur l'utilisation des véhicules privés au Canada, 1980-1996. AMDC. Final report. Reference document. Document prepared for the OEE of NRCan. Laval University, Sainte Foy. 242 pages (1998)

Merger of data from National Private Vehicle Use Survey and Manufacturer's Laboratory-tested Fuel Consumption Rates. AMDC. Final report. Document no. N98-15fa prepared for the OEE of NRCan. Laval University, Sainte Foy. 39 pages (1999)

Prediction of National Data on Private Vehicle Use in Canada, 1980-1996. AMDC. Interim report. Document no. N98-19ia prepared for the OEE of NRCan. Laval University, Sainte Foy. 263 pages (2000)

Boucher, Nathalie and Bonin, Sylvie: Seasonality and Extrapolation of NaPVUS Data. Part 1: Complete National Data Series for the Period 1980-1996. AMDC. Interim report. Document no. N99-20f1a prepared for the OEE of NRCan. Laval University, Sainte Foy. 82 pages (2000)