ABSTRACT

MAIA, LUIZ FLÁVIO. Revealed Preference and Time series Analyses of U.S. Macroeconomic Aggregates. (Under the direction of John J. Seater).

This research extends the literature on the revealed preference analysis of macroeconomic aggregates in multiple ways. The relevance of recent methodological changes in data construction is our first topic, as Varian's (1982, 1983) nonparametric tests are run on U.S. consumption series built under NIPA's old and new methods. The results indicate that previous conclusions on the overall GARP-consistency of data and on weak separability of particular aggregates are affected by the methodological changes in data. Additionally, test results are observed to be sensitive to the adoption of series at different frequencies. The issue of temporal aggregation is examined in two ways. We initially show that those changes do not seem to have significantly altered the univariate time-series properties of aggregates or previous conclusions about the impacts of temporal aggregation on those properties; therefore, the aggregation of economic flows into annual figures is once more found to involve significant losses of information about the dynamic behavior of higher-frequency data. The power of the GARP test in datasets of different frequencies is then investigated from analytical and empirical standpoints. Time aggregation is found to reduce the power of the GARP test. Finally, we apply Varian's tools to study for the first time a dataset including the value of nonmarket services produced inside the household. The modification involves a more detailed picture of consumers' allocation of time, alternatively a source of utility (leisure) or a resource in household production. We observe that the changing number of hours spent on average in household production – due to the increasing participation of women in the civilian labor force over recent decades – can be characterized as a rational decision made by the representative agent in a standard utility maximization model.

Keywords: Consumption expenditures, NIPA data, revealed preference analysis, GARP, weak separability, temporal aggregation, household production.

Revealed Preference and Time series Analyses of U.S. Macroeconomic Aggregates

by

Luiz F. Maia

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy in

ECONOMICS

Raleigh, 2004

APPROVED BY:

_____	_____
Professor John J. Seater Chair of Advisory Committee	Professor Douglas K. Pearce
_____	_____
Professor Walter N. Thurman	Professor John S. Lapp

Technical Consultant: Professor Adrian R. Fleissig

To my parents Ademilde and Luiz Flávio

and to my beloved wife Rita.

Biography

Luiz Flávio Arreguy Maia Filho was born in Belo Horizonte, Brazil, to parents Luiz Flávio Arreguy Maia and Ademilde Fonseca Arreguy Maia on August 15, 1972. He earned his Bachelor's degree in Economics in February of 1995 at the Federal University of Minas Gerais (UFMG), and his Master's of Science degree in Economics in December of 1997 at University of Brasília (UnB), both in Brazil. He lectured at UnB and worked for a year as a consultant for the Ministry of Finance of Brazil, before beginning his Ph.D. studies at North Carolina State University in August of 1998. His thesis defense was on August 16th, 2004. After graduating, he will return to Brazil and resume his career as economics professor and consultant.

Acknowledgements

I gladly acknowledge CAPES-Brazilian Government's financial support for my PhD program over the period 1998-2002, process #1856/97. I thank Professors A. Sinan Unur, Paul Fackler, Hal Varian and Mr. Tony Hetherington (Prospero Software) for many helps in obtaining and developing codes that run GARP and weak separability tests in different programming languages. I acknowledge having benefited enormously from MOSEK ApS' student licensing of their optimization software (Mosek's MATLAB toolbox, version 3), which solves large linear programming problems very efficiently. Thanks also to Drs. Stephen Landefeld and Stephanie McCulla, who provided detailed tables used in their estimation of the value of household services. The members of my advisory committee, Professors Douglas Pearce, Walter Thurman and John Lapp provided extremely helpful comments and suggestions; I deeply appreciate their friendliness and support. Professor Adrian Fleissig served as a technical consultant for this research project, but his overall contribution was genuinely equivalent to that of a Co-Advisor. Thanks, Adrian. Professor John Seater's supervision was challenging and rewarding, but most of all a privilege for which I will always be grateful. Last, I thank all my good friends, professors and students in the Economics Department, for their support and help: Varun Kshirsagar (I owe you big time, my friend!), Claudiney Pereira, Nando Hintze, Peyton Ferrier, Erin Mixon, Victoria Mitchell, Todd McFall, Pedro Oviedo, Franz Hamann, Carolyn Smith, Robin Barnett, Areendam Chanda, Ed Erickson, David Flath, Lee Craig, Jim Easley and Steve Margolis (thanks, boss!).

Table of Contents

List of Tables

List of Figures

Introduction

Aggregations along time and across goods are mostly regarded as maintained assumptions of great importance in empirical studies of macroeconomic consumption. For more than 20 years, though, Varian's (1982, 1983) nonparametric tests of revealed preferences have been applied to consumption expenditure data as a way to reduce the arbitrariness of choices concerning aggregation across goods in empirical studies. GARP-consistency and weak separability tests are run on presumably separable subcategories of per capita consumption expenditures to confirm or reject the existence of a well-behaved macro utility function rationalizing those figures. If a set of aggregates passes both tests, empirical researchers can set out maximization models for those goods only – conditional on total expenditure in that consumption category – and save degrees of freedom in the estimation of preference parameters from the optimality conditions.

Three of the four chapters in this dissertation (chapters 1, 3 and 4) contain direct contributions to this line of research. They either improve upon previous descriptions of the representative consumer's choices or study Varian's nonparametric approach itself, with the issue of temporal aggregation being systematically investigated from analytical and empirical perspectives for the first time. Chapter 2 uses a different approach (time series analysis) to reassess previous findings concerning particular aspects of temporal aggregation. Its motivation, nevertheless, is clearly connected to our discussion in chapter 1, in the sense that the relevance of methodological changes in data construction is investigated in both cases.

Together, the four parts of this work attempt to provide useful guidance for future empirical macroeconomic studies and also for general applications of GARP and weak separability tests on both macro and micro data.

Chapter 1

Revealed preference analysis of U.S. aggregate consumption: Do revised NIPA data provide new support for the representative agent approach?

1.1 Introduction

Since its 10th Comprehensive Revision, real values in NIPA tables have been estimated from Fisher indices and presented mostly in chained dollars, rather than from the traditional Laspeyres price/quantity indices. Landefeld and Parker (1997) argue that the new measure accounts for changes in relative prices over time, "*(…)thereby eliminating a major source of bias in the previously featured fixed-weighted, or Laspeyres, measures of real output and prices*". One can interpret such a methodological change as a significant reduction in data measurement error. Nevertheless, the impact of this improvement – along with two other relevant methodological changes, to be discussed later – on prior empirical findings concerning GARP-consistency and weak separability of U.S. consumption expenditures have not yet been studied, constituting our main goal in this chapter.

Perhaps the most appropriate reference for comparison of any new findings, Fleissig, Hall and Seater (2000), hereafter FHS, applied the aforementioned nonparametric tests to U.S. consumption data at multiple frequencies, starting from disaggregated categories of U.S. consumption expenditure. Some of the previous papers had relied on major consumption categories calculated from the simple sum of more disaggregate data[1]; FHS objected that such a procedure presupposes an amount of separability that might not be consistent with the data. They exploited two important advantages of Varian's nonparametric framework: its power to handle a large number of goods – actually, consumption subcategories – and the fact that no presumed aggregation structure is required. They found that over GARP-consistent samples: (i) the aggregate utility function was separable in nondurables and services, but not in durables; (ii) durables could be aggregated up to two or three major categories, but not all the way up to one overall aggregate; (iii) methods of aggregation must be consistent with the structure of separability; (iv) results were sensitive to the adoption of specific data frequencies, with monthly data containing the largest number of GARP violations[2]. In a subsequent paper, Fleissig, Gallant and Seater (2000) showed that accounting for the first three of those aspects had substantial impacts on the estimated preference parameters from Euler equations and also on the overall (non)rejection of intertemporal consumption models.

Other papers have used aggregate data to study, among other things, the consumption-leisure separability (Swofford and Whitney, 1987, 1988; Drake, 1997). Swofford and Whitney (1987) – henceforth SW – studied 15 years of U.S. quarterly data and found that consumption goods and leisure were weakly separable from monetary assets in the representative utility function. Adopting data on the three major aggregates of consumption expenditures (durables, nondurables and services), they observed that

[1]Notoriously, Swofford and Whitney (1987, 1988). Drake (1997) reports having checked the robustness of their findings by running the tests also in disaggregate categories of UK consumption expenditures. Other studies focused on more limited subsets of consumption data (Varian, 1982; Bronars, 1987; Alston and Chalfant, 1988).

[2]We will investigate the impacts of temporal aggregation on the power of the GARP test from analytical and empirical standpoints in chapter 3.

consumption goods alone were not weakly separable from two other presumed sources of utility, such as leisure and monetary assets. Swofford and Whitney (1988) expanded their analysis to compare results from quarterly and annual data and to consider evidence supporting partial adjustment models for the allocation of monetary assets.

Besides discussing and adopting data that was subject to important methodological changes, this chapter improves upon the available evidence in multiple ways. First, we expand FHS's analysis of (disaggregated) consumption subcategories and reevaluate consumption-leisure weak separability. In order to study the weak separability of alternative aggregates, we implement a fairly recent algorithm developed by Fleissig and Whitney (2003) that improves upon the method used in Varian's original software[3]. Finally, we will benefit from datasets with samples longer than most of the previously studied ones, including series at different frequencies.

We must acknowledge two sources of criticism to the approach just described. The first one is due to data choices for the consideration of leisure/labor optimal decisions of a representative agent. Some economists are skeptical about the usually adopted proxies for the price of leisure, measured as the opportunity cost of time not committed to work[4]. Following both SW and Drake (1997), we will once more adopt a wage rate to measure the price of leisure hours[5].

The second concern comes from the nature of the nonparametric tests of GARP consistency developed by Varian (1982). As it is nonstochastic, the test can reject the existence of a well-behaved representative utility function rationalizing the data due to a single observation deviating from the expected maximizing behavior. Recently,

[3]Virtually all previous results in the literature relied on Varian's software, either in its PASCAL parallel-computing version or in the one for PC's – which limits the numbers of goods/observations. Building upon Anan Usur's (Cornell University) MATLAB codes to test GARP, we implemented all tests for this research project in that programming environment – codes available upon request. To solve the nontrivial linear programming problem involved in Fleissig and Whitney's (2003), MATLAB's built-in algorithm is very inefficient, and we had to use MOSEK's toolbox (student version), which is accessed from MATLAB.

[4]I thank Adrian Fleissig and John Seater for having pointed this out, as such an issue is seldom explicitly considered in this particular literature; not tackled in this research project, addressing the issue on both theoretical and empirical grounds seems to be a promising direction for future extensions.

[5]Alternatively, SW considered Barnett's (1979) shadow price of leisure, finding similar results, though.

however, Varian (1996) discussed and implemented an algorithm to assess goodness-of-fit of consumption data, as originally suggested by Afriat (1972). Even though this additional test is itself subject to criticism, due to its relatively low power against "naive" alternative hypotheses (Sippel, 1996), it will provide some information about how large violations are and, additionally, a basis for comparisons of data at different frequencies.

The remainder of the chapter is divided into four additional sections. In section 1.2 we briefly discuss the tests and procedures that will be used throughout this dissertation. Section 1.3 provides information on the methodological changes in data[6]. Then we present and compare our results to previous findings in section 1.4, and the last section concludes.

1.2 Nonparametric tests of GARP consistency and Weak Separability

Rather than repeating all definitions, theorems and proofs for the tests that will be used later, the current section is meant to assist the reader understanding the procedures and interpreting the results. Special attention will be drawn to the efficiency test, less often discussed in the literature, and to improvements on Varian's weak separability test recently suggested by Fleissig and Whitney (2003).

1.2.1 GARP consistency and Consumption Efficiency

In Varian's (1996) implementation of Afriat's (1972) efficiency index, the main goal is to check how close observed consumption choices come to satisfying GARP. Intuitively, if a few observations out of a series of choices violate GARP, there may be a

[6]Appendix 1 presents further details on data sources and manipulations.

small perturbation of the budget constraints that makes all data consistent with optimizing behavior[7]. In this subsection we describe both the traditional GARP-consistency test developed by Varian (1982) and the essentials of an algorithm to measure goodness-of-fit of consumption data to GARP (Varian, 1996), showing how the former can be seen as a special (stricter) case of the latter.

Consider a standard utility-maximizing consumer that chooses a vector of goods **x**, facing a vector of corresponding prices **p** and total income **m** – superscripts denoting specific observations. Let $\mathbf{p}^t$ be the vector of current prices when a choice $\mathbf{x}^t$ is made; we say that $\mathbf{x}^t$ is directly revealed preferred to an alternative $\mathbf{x}^s$ if and only if $\mathbf{x}^t$ is purchased when $\mathbf{x}^s$ is also affordable:

Directly Revealed Preference (DRP): $\mathbf{x}^t \mathbf{R}^D \mathbf{x}^s \iff \mathbf{p}^t\mathbf{x}^t \geq \mathbf{p}^t\mathbf{x}^s$

Intuitively, we cannot say that $\mathbf{x}^t$ is revealed-preferred to any other bundle that was not actually affordable when the choice was made. Two extensions to this basic relation are relevant. First, one can define *Strict Revealed Preference Relation* (R^S) in the same lines as DRP, but changing the inequality signal to ">" (strictly greater than). Additionally, we label the revealed preference relation R as the transitive closure of the relation R^D; that is, $\mathbf{x}^t R \mathbf{x}^z$ if and only if there is some chain of observations ($\mathbf{x}^t$, $\mathbf{x}^u$, $\mathbf{x}^v$,…, $\mathbf{x}^z$) such that $\mathbf{x}^t R^D \mathbf{x}^u$, $\mathbf{x}^u R^D \mathbf{x}^v$,…, $\mathbf{x}^y R^D \mathbf{x}^z$.

Assume now that $\mathbf{x}^t$ is directly revealed preferred to $\mathbf{x}^s$, as in our definition of DRP; also, as goods are sold later at a new set of prices, $\mathbf{p}^s$, $\mathbf{x}^s$ is chosen; this behavior is not consistent with the utility maximization model if, at the new prices $\mathbf{p}^s$, $\mathbf{x}^s$ turns out to be chosen and $\mathbf{x}^t$ is also affordable. In other words, if $\mathbf{x}^s$ and $\mathbf{x}^t$ were affordable at both occasions, the consumer should have revealed the same preference between them and made the same choice. Therefore, the pair of observed choices is only consistent with the maximization model if $\mathbf{x}^t$ is not affordable later, as $\mathbf{x}^s$ is chosen. GARP is defined as:

[7]Our presentation follows closely the discussion in Varian (1996).

GARP: if $x^t R x^s$ at current prices p^t but x^s is chosen at a different set of prices p^s, then it is not the case that $p^s x^s > p^s x^t$, i.e., it is not true that $x^s R^S x^t$.

The definition above contains all important elements of Varian's (1982, 1983) nonparametric test, which simply verifies the occurrence of GARP violations in a series of consumption choices. Equivalently, one can say that the test verifies whether the consumer's preference over a set of observed choices remains the same over the sample.

As mentioned before, one may also want to consider how close to satisfying GARP the consumer choices are, given violations occur. For that we must restate the preference relations, allowing small deviations from strict maximizing behavior; we now define Direct Revealed Preference (DRP) at specific levels of efficiency:

DRP at the efficiency level e (R^{DE}): for $0 < e \leq 1$, $x^t R^{DE} x^s$ <=> $e p^t x^t \geq p^t x^s$

We say that $\mathbf{x^t}$ is directly revealed preferred to some hypothetical alternative $\mathbf{x^s}$ at the efficiency level e if and only if $\mathbf{x^t}$ turns out to be chosen and $e\mathbf{p^t x^t} \geq \mathbf{p^t x^s}$; since $e \leq 1$, the observed choice $(\mathbf{p^t}, \mathbf{x^t})$ is now considered revealed-preferred only to alternatives that could have been afforded with a fraction (e) of the actual expenditure. As a consequence, any choice actually made will reveal a consumer's preferences over a potentially smaller set of alternatives (except if e=1, which implies the original direct revealed preference relation), much cheaper ones. One can immediately suspect that smaller efficiency levels will make GARP violations less likely. Varian's (1996) economic interpretation for this efficiency level is that a part of the consumer's income (1-e) could have been "wasted" or not optimally allocated. Another interpretation is that one can only conclude about a consumer's preference for a bundle over alternatives when those alternatives are significantly less expensive than the original choice.[8] If the consumer declines much cheaper alternatives, it may be most certainly due to the specifics of his tastes over those choices – and less likely due to a certain indifference between very close alternatives

involving insignificantly different expenditures. In any case, we can use this R^{DE} relation to compute its transitive closure R_e as before and ultimately check for $GARP_e$, that is, GARP violations at specific levels of efficiency[9]:

GARP at the efficiency level "e"($GARP_e$): if $x^t R_e x^s$, then $ep^s x^s < p^s x^t$.

To say that a specific set of choices satisfies $GARP_e$ at 98% efficiency (e=0.98) means that we have to relax every budget constraint up to 2% so that no pair of choices can be revealed inconsistent. In other words, we deliberately disregard eventual inconsistencies that are considered small enough and, therefore, insignificant. Only choices that indicate significant losses to the consumer's welfare will be treated as violations in this case; we use the term "loss" in the sense that the consumer could have acquired an alternative bundle revealed at least as good as his actual choice, spending (up to 2%) less. Varian (1996) then suggests a simple optimization algorithm (binary search) to compute the largest value of e such that no violation of $GARP_e$ occurs; this maximum value is referred to as Consumption Efficiency Index[10].

Afriat first suggested the efficiency measure discussed above more than 30 years ago, but its implementation was impractical for larger data sets until recently. Computational advances and their very incorporation into economists' set of tools revealed otherwise.

A few words of caution are due, however, as some papers have found that GARP tests at efficiency levels different than e=1 have low power against alternative hypotheses, such as purely random behavior. Bronars (1987) simulated purely random demand data and evaluated GARP-consistency of both aggregate and per-capita consumption; his results indicated that the power of the nonparametric test against the

---

[8]In a dynamic framework, short-run inconsistencies could be attributable to the dissemination of information or habit persistence, aspects not present in the underlying (basic) utility-maximization model.

[9]Varian assumes that a choice is always directly revealed preferred to itself (not required, though).

[10]Afriat originally called this number "critical cost efficiency index"; other papers refer to this "e" maximum as Afriat Efficiency index/level.

alternative hypothesis was over 90% with per capita data only[11]. In other words, the original GARP test (at 100% efficiency) actually rejected the optimizing-behavior hypothesis quite frequently with randomly generated per capita data. On the other hand, Sippel (1996) found that the relaxation of full efficiency inherent to the consumption efficiency measure aforementioned could be quite misleading; "at the 'prominent' efficiency level of 95% almost 90% of simulated random-demand data passed the consistency test". Concluding, researchers should not believe that a high level of consumption efficiency guarantees near-optimizing behavior; *au contraire*, even a 95% efficiency level should raise suspicion about the validity of the null hypothesis. In a recent paper, Drake (1997) implemented the overall efficiency measure discussed above – referring to it as "Goodness-of-Fit" – to a relatively small sample of UK data[12]. His results were sensitive to the scaling of aggregate data, per capita or per household; nevertheless, efficiency levels of per household data were never below 99.82%, whereas per capita data revealed efficiency levels as low as 97.32%.

Following Drake (1997) and rather than drawing conclusions on the absolute values of the results, comparing efficiency levels of consumption data at various frequencies (quarterly and annual) and over different samples seems to be an appropriate approach.

1.2.2 Interpreting and testing weak separability

The nonparametric test of weak separability was originally developed and implemented by Varian (1983); nevertheless, the results that will be presented in the coming sections actually rely on Fleissig and Whitney's (2003) improved algorithm to search for numbers satisfying the so called "Afriat inequalities", as we discuss next[13].

[11]Bronars' (1987) and other simulation exercises will be considered in details in chapter 3.
[12]Approximately seven years of quarterly data, 1986-1993.
[13]The last three tables at the end of this chapter will show results using Varian's software, for comparison.

Again, one can avoid restating theorems and proofs, but it is useful to define and illustrate weakly separable preferences before we discuss the test and distinguish between necessary and sufficient conditions of the test.

Using FHS's notation, suppose that a vector of k goods is partitioned into two subsets, **a** and **b**, where **a**=(x_1, $x_{2,\ldots}$,x_m) and **b**=(x_{m+1}, $x_{m+2}\ldots$,x_k); a utility function U(**x**) is weakly separable in b-goods if there exist a subutility function v(**b**) and a macro function u*[**a**,v(**b**)] which is continuous and monotonically strictly increasing in v(b), such that U(**a**,**b**)≡ u*[**a**,v(**b**)]. Two facts must be remembered: first, if a utility function is weakly separable in **b** goods, it means that the marginal rate of substitution between any two of those goods is independent of the "**a**" goods; second, separability in "**b**" goods does not imply separability in "**a**" goods[14].

The most common functional forms defining preferences in economics are weakly separable, including the Cobb-Douglas and the CES (Constant Elasticity of Substitution) specifications. Blackorby et al. (1998) provide the following example of nonseparable functional format, which was later adopted in Fleissig and Whitney's (2003) simulation exercises:

$$U(x_1,x_2,x_3,x_4) = x_1^{1/3}x_3^{1/3}x_4^{1/3} + x_2^{1/2}x_3^{1/4}x_4^{1/4}$$

Defining $v(x_3,x_4) = x_3^{1/3}x_4^{1/3}$ and $u^*(x_1,x_2,v) = x_1^{1/3}v + x_2^{1/2}v^{3/4}$, notice that

$$U(x_1,x_2,x_3,x_4) = u^*[x_1,x_2,v(x_3,x_4)]$$

Further, one can define $U_i(x)$ as the first derivative of the utility function $U(x)$ with respect to commodity "i" and use tedious algebra to show that $\frac{\partial}{\partial x_1}\left(\frac{U_3(x_1,x_2,x_3,x_4)}{U_4(x_1,x_2,x_3,x_4)}\right)=0$ and that $\frac{\partial}{\partial x_3}\left(\frac{U_1(x_1,x_2,x_3,x_4)}{U_2(x_1,x_2,x_3,x_4)}\right)\neq 0$; in other words, the marginal rate

[14]See Pollak (1971) for a good discussion on separability concepts and their main implications.

of substitution between x_3 and x_4 does not depend on the levels of consumption of goods outside that "branch", whereas the same cannot be said about x_1 and x_2. The function is weakly separable in the goods x_3 and x_4 but not in x_1 and x_2.

A necessary condition for weak separability is that the supposedly separable subset of data must first pass the GARP-consistency test, which can indicate the existence of a well-behaved subutility function v(.); contingent on such a result, we can proceed to check a sufficient condition of weak separability. This second part of the test relies on the existence of numbers satisfying a series of inequalities involving prices, quantities and expenditures on each supposedly separable subset of goods (throughout a finite data sample). Those numbers, interpreted as each group's quantity and price indexes, must also conform to the GARP-consistency of the overall set of goods in the macro utility function. Varian's (1983) separability theorem states that the existence of numbers satisfying the "Afriat inequalities" and that overall GARP-consistency is equivalent to the existence of a well-behaved weakly separable utility function rationalizing the data.

Also based on Varian's theorem, Fleissig and Whitney (2003) recently developed a more efficient algorithm to search for the Afriat numbers, which starts from superlative group indexes and searches for the smallest necessary deviations from them (if any) so that Afriat inequalities are satisfied. The authors were motivated by results such as Barnett and Choi's (1989), which reported having used Varian's NONPAR software and failed to obtain sufficiency in simulated Cobb-Douglas data. The advantages of this algorithm are the fact that it can be implemented in PC's – with relatively low computational costs, if compared to the use of supercomputers in Swofford and Whitney (1994) – and the potential avoidance of inconclusive results, as we explain next.

As one investigates alternative separability structures, the results indicate which of them pass the tests for necessary, necessary and sufficient or none of the conditions. If a disaggregated set of goods passes both conditions for weak separability in a representative utility function, their later consolidation (aggregation) in the empirical analyses of alternative aggregator functions is not only convenient, but actually a

theoretically valid procedure[15]. Passing none of the them leads to a conclusive rejection of some separability structure. However, passing only the necessary condition means that the algorithm used to search for numbers satisfying the aforementioned Afriat inequalities failed to find them. The result is considered inconclusive because the existence of such numbers is not definitely rejected. In this respect, a more complete/efficient search for Afriat numbers will potentially reduce the number of inconclusive results[16].

As a final note in this section, we acknowledge that there are always maintained hypotheses on the testing of separability structures; the fact that our analysis does not include liquidity services from money holdings[17], for example, can be interpreted as a maintained assumption that consumption goods and leisure are weakly separable from those services, as well as from any possibly excluded source of utility[18].

[15]Testing those separability structures does not require aggregation; however, Fleissig, Gallant and Seater (2000) showed that the choice of aggregation method might significantly affect the estimation of preference parameters of models assuming specific separability structures.

[16]See Fleissig and Whitney (2003) for a thorough discussion of this improved version of Varian's test.

[17]We obviously refer here to Money-in-Utility-Function models, as in Barnett, Fisher and Serletis (1992), Fisher and Fleissig (1997) and Holman (1998).

[18]In chapter 4 we will explicitly consider the utility derived from the consumption of nonmarket household production, as estimated by Landefeld and McCulla (2000).

1.3 Methodological changes and their relevance

1.3.1 Data sources and methods

The two main sources of data used here are BEA's NIPA and Fixed Asset tables, for seasonally adjusted personal consumption expenditures and the depreciation/stock of consumer durables, respectively. The largest common sample of all series goes from 1964 to 1990[19], but subsets of data range from 1929 to 2000. A comprehensive explanation of all data manipulations is presented in Appendix 1. The focus in this section is on the methodological changes in data construction, which were essentially three: (i)Substitution of fixed-weight price and quantity indices by chain-type ones; (ii)adoption of new assumptions in the estimation of depreciation rates for durable goods; and (iii) the breaking of expenditures on Services into 6 rather than 5 subcategories, as expenditures on recreational services became an individual subcategory separated from the residual "other services". To investigate the very existence of weakly separable representative utility functions on alternative consumption aggregates, we adopted not the major categories of real personal expenditures but their components, following FHS; sets of 14 components of consumption expenditures – at different frequencies, under new and old methodologies – were compiled[20]:

Durables, 3 components: (D_1) Motor vehicles and parts,
(D_2) furniture and household equipment and
(D_3) other durables;

[19]The most relevant limitations come from labor data, with series going back only to 1964, and from all series built under the old methodology, which were dropped from NIPA tables.

[20]FHS works with the same level of aggregation for monthly and quarterly data, but a larger set of subcategories for annual data. Our approach is to start from the same set of categories to make the results from data at all frequencies more directly comparable. Once more, notice that under Nipa's old methodology there will be one less service subcategory and, therefore, only 13 consumption goods.

Nondurables, 5 components: (ND_1) Food, (ND_2) clothing and shoes, (ND_3) gasoline and oil, (ND_4) fuel oil and coal and (ND_5) other nondurables;

Services, 6 components: (S_1)Housing, (S_2)household operations, (S_3) transportation, (S_4) medical care, (S_5) recreation and (S_6) other services.

Throughout this research project we deliberately avoid the use of GARP and weak separability tests on monthly data, due to the well known fact that those figures are particularly subject to measurement problems; Wilcox (1992) pointed out the two most critical sources of imperfections: (i) monthly total retail-sales figures are estimated from samples, therefore subject to sampling errors; and (ii) product composition of retail sales is not known at the monthly frequency. In fact, it is assumed that the composition of sales within each category of stores is fixed throughout quarters. The task of constructing monthly figures is also somewhat complicated by the fact that many retailers do not tabulate their sales by calendar months, demanding further adjustments. His main conclusion is that published data – specially at the monthly frequency – cannot automatically be assumed to correspond exactly to their theoretical analogues. Researchers using high frequency data should introduce an explicit model for the sampling error, as in Bell and Hillmer (1990)[21].

As for leisure prices and quantities, we essentially followed the procedures in SW, which also assumes a 10-hour daily fixed allocation of time for sleeping and eating (see appendix for more details). In fact, Swofford and Whitney (1988, 1994), Drake (1997) and Drake et al. (2003) adopted the same fixed amount of nonmarket hours per day; Mankiw et al. (1985) also assumed a daily fixed amount of time (only 8 hours, though) allocated neither to work nor leisure. Also standard in this literature, the opportunity cost of time is proxied by the wage rate, measured as the seasonally adjusted

[21]FHS studied GARP consistency and weak separability of monthly data (not adjusting for sampling errors), along with quarterly and annual figures; they reported hundreds of GARP violations over their whole sample (1959-1990); the largest subsample of GARP-consistent monthly data covered 20 years (1970:05-1990:12).

average hourly earnings of production workers – average hourly earnings of production or nonsupervisory workers on private nonfarm payrolls. Annual (and quarterly) data are calculated as the average of original monthly data over periods. The labor data comes from the Bureau of Labor Statistics[22].

1.3.2 NIPA's adoption of superlative indices

Recent articles by Triplett (1992) and Rossiter (2000) explain the advantages of an important methodological change in NIPA tables: the substitution of price/quantity indices. We start this section summarizing those articles, so that we can better understand (later) the impact of this change on the evaluation of GARP consistency/weak separability of aggregates.

The traditional use of fixed-price indices in the calculation of NIPA figures has been subject to criticism for some time[23], but only recently did BEA opt for its replacement. Its main problem is the occurrence of what has been known as "substitution bias". Among other uses, price indices are expected to measure changing costs of a constant standard of living, while quantity indices measure changes in that standard of living. Real values in NIPA tables have been calculated from a fixed-weight price index known as Laspeyres Index. They presume a constant set of goods and services, actually acquired in a base-year, representing the referential standard of living. Two problems emerge from such an assumption in the actual decomposition of observed changes in expenditures into price and quantity components: first, consumers may obtain the same standard of living from different sets of goods and services, but actual changes in the

[22]The particular series was identified in the BLS website with the code/number EES00500006. The following description is extracted from BLS's Handbook of Methods (also available online): "Average hourly earnings series, derived by dividing gross payrolls by total hours, reflect the actual earnings of workers, including premium pay. They differ from wage rates, which are the amounts stipulated for a given unit of work or time. Average hourly earnings do not represent total labor costs per hour for the employer, because they exclude retroactive payments and irregular bonuses, employee benefits, and the employer's share of payroll taxes. Earnings for those employees not included in the production worker or nonsupervisory categories are not reflected in the estimates."

composition of these bundles are not captured with fixed-weight indices; second, consumers tend to substitute away from those goods whose prices rise fast, towards the ones whose prices rise slower or fall. Consequently, measuring price changes in periods before some base-year will tend to give too much weight to goods whose prices have risen fast and too little to the ones whose prices have fallen; one can expect that if overall price changes over some period are overestimated, changes in real aggregates (or in the standard of living) calculated from the same index will tend to be underestimated. Periodically, the very much necessary updating of the referential standard of living ends up making those issues even more evident, as historical figures change significantly with the periodical adoption of a new base-year.

The "substitution bias" can be measured precisely as the difference between fixed-weight price indices and price indices that somehow account for that sort of substitution. Diewert (1976) showed that there existed relatively simple ways to approximate an ideal (unbiased) theoretical measure for the cost-of-living. Twenty years later, one of those so called "superlative" price indices – known as the Fisher Ideal index – has finally been adopted in the computation of NIPA aggregates.

Fisher Ideal indices (quantity and price) are defined as the geometric average of two other indices, Laspeyres and Paasche indices. The first, as described above, has the convenience of relying on referential quantities/prices from a single period or base-year; Paasche indices, on the other hand, assume current sets of prices or quantities to estimate changes between a given period and some base-year, which makes them more costly and computationally burdensome. Rossiter (2000) discusses the differences between those indices in practical evaluations of price/quantity changes and presents simple examples showing that whenever Laspeyres indices tend to overestimate price changes, Paasche indices will underestimate them, and *vice versa*. The choice of an average between these two indices to obtain unbiasedness is, therefore, a natural one.

---

[23]Triplett (1992) cites two major studies: Braithwait (1980) and Manser and MacDonald (1988).

All previous empirical studies on GARP consistency of U.S. aggregate consumption have relied on NIPA figures calculated under the traditional methodology, data availability being the fundamental constraint. To our knowledge, this is the first time Varian's nonparametric tests are implemented on NIPA data calculated under the new methodology; therefore, we must first compare analytically the differences between the indices and their particular manipulation in GARP-consistency tests. We show that adopting Laspeyres or Fisher-based aggregates does in general affect GARP-consistency and weak separability tests, for two reasons: not only do they entail distinct decompositions of changes in nominal expenditures into price and quantity movements, potentially affecting relative prices, they also imply different measures of available income at the time choices are made.

As described in the previous section, all GARP consistency tests demand vectors of prices and quantities, so that one can identify all feasible choices when one particular bundle is elected and acquired by the consumer. In past analyses of aggregate consumption, the quantities acquired by a representative consumer were "proxied" by constant real dollar expenditures on a specific category of goods, whereas fixed-weight price indices were readily admitted as the corresponding normalized prices for each category. Such assumptions naturally imply a corresponding proxy for the total income allocated to consumption (or total consumption expenditure), which may not correspond to the actually observed nominal expenditure at each period. To see this, consider a set of n nondurable goods grouped in a specific subcategory; for simplicity, we will admit a short series of only three observations, adopting the second period as a base-year for the indices. Following the traditional methodology of NIPA tables, we can express the series of observed prices as below:

$$p^{LASP} = 100 \times \begin{bmatrix} \left(\sum_{n=1}^{N} P_1^n Q_2^n / \sum_{n=1}^{N} P_2^n Q_2^n \right) \\ \left(\sum_{n=1}^{N} P_2^n Q_2^n / \sum_{n=1}^{N} P_2^n Q_2^n \right) \\ \left(\sum_{n=1}^{N} P_3^n Q_2^n / \sum_{n=1}^{N} P_2^n Q_2^n \right) \end{bmatrix} \qquad \textbf{(1.1)}$$

where P^n_t is the price of good n at time t, Q^n_t is the quantity of good n acquired at time t and $\sum$ represents the sums over n goods. Once more, the interpretation of Laspeyres price indices is straightforward: it represents a ratio of current and base-year expenditures, with the exact same set of goods being admittedly purchased at different prices. The vector of quantities are calculated once more from a Laspeyres formula (quantity index), multiplied by current expenditures in the base-period:

$$q^{LASP} = \sum_{n=1}^{N} P_2^n Q_2^n \times \begin{bmatrix} \left(\sum_{n=1}^{N} P_2^n Q_1^n / \sum_{n=1}^{N} P_2^n Q_2^n \right) \\ \left(\sum_{n=1}^{N} P_2^n Q_2^n / \sum_{n=1}^{N} P_2^n Q_2^n \right) \\ \left(\sum_{n=1}^{N} P_2^n Q_3^n / \sum_{n=1}^{N} P_2^n Q_2^n \right) \end{bmatrix} \quad \textbf{(1.2)}$$

As mentioned before, a corresponding vector of available income can be easily calculated from the product of prices and quantities (actually, the product of proxies for actual prices and quantities) in each period:

$$M^{LASP} = 100 \times \begin{bmatrix} \left(\sum_{n=1}^{N} P_1^n Q_2^n / \sum_{n=1}^{N} P_2^n Q_2^n \right) \times \sum_{n=1}^{N} P_2^n Q_1^n \\ \left(\sum_{n=1}^{N} P_2^n Q_2^n \right) \\ \left(\sum_{n=1}^{N} P_3^n Q_2^n / \sum_{n=1}^{N} P_2^n Q_2^n \right) \times \sum_{n=1}^{N} P_2^n Q_3^n \end{bmatrix} \quad \textbf{(1.3)}$$

Notice that we have no reason to believe that the first element of the income vector M will equal the total (current) expenditures actually made in that period ($\sum P_1 Q_1$). In this sense, the proxy for available income is somewhat counterintuitive, if not inappropriate.

We can now verify that with Fisher indices, the proxied available income in each period will hold a close correspondence to the consumption expenditure at current dollars. Since Fisher Ideal indices are essentially geometric averages of Laspeyres and Paasche indices, the formulas are slightly more complicated:

$$p^{FISH} = 100 \times \begin{bmatrix} \left(\left(\sum_{n=1}^{N} P_1^n Q_2^n / \sum_{n=1}^{N} P_2^n Q_2^n \right) \times \left(\sum_{n=1}^{N} P_1^n Q_1^n / \sum_{n=1}^{N} P_2^n Q_1^n \right) \right)^{0.5} \\ \left(\left(\sum_{n=1}^{N} P_2^n Q_2^n / \sum_{n=1}^{N} P_2^n Q_2^n \right) \times \left(\sum_{n=1}^{N} P_2^n Q_2^n / \sum_{n=1}^{N} P_2^n Q_2^n \right) \right)^{0.5} \\ \left(\left(\sum_{n=1}^{N} P_3^n Q_2^n / \sum_{n=1}^{N} P_2^n Q_2^n \right) \times \left(\sum_{n=1}^{N} P_3^n Q_3^n / \sum_{n=1}^{N} P_2^n Q_3^n \right) \right)^{0.5} \end{bmatrix} \tag{1.4}$$

$$q^{FISH} = \sum_{n=1}^{N} P_2^n Q_2^n \times \begin{bmatrix} \left(\left(\sum_{n=1}^{N} P_2^n Q_1^n / \sum_{n=1}^{N} P_2^n Q_2^n \right) \times \left(\sum_{n=1}^{N} P_1^n Q_1^n / \sum_{n=1}^{N} P_1^n Q_2^n \right) \right)^{0.5} \\ \left(\left(\sum_{n=1}^{N} P_2^n Q_2^n / \sum_{n=1}^{N} P_2^n Q_2^n \right) \times \left(\sum_{n=1}^{N} P_2^n Q_2^n / \sum_{n=1}^{N} P_2^n Q_2^n \right) \right)^{0.5} \\ \left(\left(\sum_{n=1}^{N} P_2^n Q_3^n / \sum_{n=1}^{N} P_2^n Q_2^n \right) \times \left(\sum_{n=1}^{N} P_3^n Q_3^n / \sum_{n=1}^{N} P_3^n Q_2^n \right) \right)^{0.5} \end{bmatrix} \tag{1.5}$$

The product of price and quantity proxies at each period will result in the vector of proxies for available income; the proxy for available income in this case is simply a multiple of the actual consumption expenditure:

$$M^{FISH} = 100 \times \begin{bmatrix} \sum_{n=1}^{N} P_1^n Q_1^n \\ \sum_{n=1}^{N} P_2^n Q_2^n \\ \sum_{n=1}^{N} P_3^n Q_3^n \end{bmatrix} \tag{1.6}$$

We have shown that the adoption of different price indices results in clearly distinct proxies for actually chosen quantities and prices, as well as for available income; consequently, the set of "feasible but not chosen" bundles at some specific period may not be the same under old and new methodologies[24].

[24]Notice that the measures of consumption efficiency defined earlier are more likely to reach different conclusions if data contain the characteristic that makes disparities between different indices large, or in

To illustrate the impacts of methodological changes on the evolution of prices and quantities, we plotted annual figures of a few subcategories of per capita consumption expenditures calculated under both methods. Figures 1.1 and 1.2 show real per capita expenditures and prices indices of food under both methodologies, respectively. A careful observation of these diagrams reveals that, despite the different year basis, the adoption of different methods implies very distinct decomposition of changes in nominal expenditure into price and quantity movements: the fixed-weight or Lapeyres method (continuous lines in both diagrams) tends to overestimate price changes as a component of the changes in nominal expenditure, whereas the overall change in real quantity is underestimated.

One can also see that the choice of index affects the observed changes in relative prices over time. Figure 1.3 plots ratios of price indices for food (pnd1) and household operations (ps2), an example in which relative prices evolve very distinctly under old and new methodologies. Referring back to our description of Varian's GARP consistency test, remember that each period's observed consumption bundle is checked in the search for affordable but not chosen bundles at all times. Biased relative prices can and typically will affect the set of affordable but not chosen bundles at each time, which ultimately compromise the reliability of GARP test conclusions[25].

The adoption of Fisher indices is certainly an improvement to NIPA's calculation of aggregate price/quantity changes, but some practical problems have arisen from their uses by researchers/analysts who are not familiar with the new and slightly more complex methodology. The most common of these problems has to do with the lack of additiveness of real aggregates calculated under the new methodology. Because Fisher indices rely on geometric averaging of price/quantities from distinct periods, real

---

other words, the presence of large substitution biases. As pointed out by Manser and McDonald (1988), two main factors are a)the occurrence of significant changes in relative prices among disaggregated components and b)higher degrees of substitutability between commodities. Those will typically occur as one considers observations fairly distant from the base-year. We will return to this discussion as we analyze and compare our results with previous findings in the literature.

aggregates calculated under this methodology will not typically equal the sum of their components. The "residuals" can be attributed to changes in relative prices between the current period and the base-year; therefore, they will tend to be smaller for periods close to the base-year and larger for distant ones[26].

However, the lack of additivity of real aggregates calculated under the new methodology is not a concern in our analysis: GARP tests rely on the sum of nominal expenditures – the product of prices and real quantities – to evaluate whether alternative consumption bundles are affordable at each period; even with the new methodology, the sum of expenditures in components does equal total expenditure in an aggregate. In other words, there's additivity in expenditures.

Emphasizing this point once more, the reader should keep in mind that the tests to be conducted here never require the summation of real aggregates, which with the new data methodology would mean adding up oranges and bananas. All examinations of revealed preference orderings are made on nominal terms. The GARP test evaluates nominal expenditures involved in the purchase of bundles at different (current) prices; it does not involve comparing the proxies for quantities (real expenditures) directly.

---

[25]Figure 3 shows an extreme case, in which observed relative prices are very significantly affected by the choice of index; however, as pointed out by Adrian Fleissig, even small differences in relative prices can lead us to conclude erroneously that consumer's choice are (are not) inconsistent with GARP.

[26]More precisely, real aggregates differ from the sum of their components because changes in expenditure on each of the components are decomposed in particular contributions from quantity and price movements. Ironically, such is the source of both the strength (unbiased) and what some may consider the weakness (lack of additivity) of Fisher-index aggregates.

1.3.3 New depreciation profiles and estimates for user costs of durables[27]

BEA's data on depreciation of consumer durables have also been revised significantly. Katz and Herman's (1997) *Survey of Current Business* article about those changes presents interesting points that will help us understand their possible implications to our analysis. It must be noticed first that BEA estimates and publishes only annual depreciation of fixed assets, measured as the "decline in value due to wear and tear, obsolescence, accidental damage and aging". Figures are presented both in current values and in real chained-dollars. Under both old and new methodologies, estimates of net stocks and depreciation are derived using the familiar perpetual inventory method: the net stock is calculated as the difference between cumulative values of past gross investment and depreciation.

The most important difference between new and old methodologies comes from the assumptions underlying the calculation of depreciation rates. The new method makes use of empirical evidence on prices of used equipment and structures in resale markets, which have shown that depreciation for most types of assets approximates a geometric pattern. Table 1.1 shows a simple example comparing new and old depreciation profiles. Previously, BEA assumed what is referred to as "Straight-Line Depreciation Profile": assuming a service life of 10 years for some equipment, it would lose an equal percentage of its original value each year. Calculated on a period-by-period basis, the depreciation rate (defined simply as the ratio between current depreciation and the sum of current depreciation and net stock) would be actually increasing as time passes. Under the new methodology – the Geometric Pattern Depreciation Profile – the depreciation rate is constant, but often larger than simply the reciprocal of the number of service-years (1/L): it is multiplied by a "declining-balance" index, this one estimated from empirical studies on similar classes of assets.

[27]The changes discussed in this section only affect our analysis of annual data, since BEA only publishes annual figures on depreciation; refer to Appendix 1 for details on how quarterly depreciation rates were calculated.

This declining balance index is generally larger than 1 for equipment and therefore, for most consumer durables. Under the new methodology, the actual depreciation rate tends to become larger in the first years of the equipment's service life and shorter later, if compared to the old methodology's depreciation pattern. In our simple example, the (typical) declining-balance index is 1.65 and the depreciation rate becomes larger for the first 4 years of this hypothetical equipment's service life and shorter thereafter. The calculation of annual depreciation of assets does consider that pieces of equipment of different vintages may have different depreciation profiles[28]. Thus, final figures express a weighted average of depreciation flows from assets of different vintage.

Precisely as in FHS and SW, consumers are assumed to obtain utility from services proportional to the stocks of durables they hold; the price of those services is calculate as the user cost of holding those assets over each period[29]:

$$uc_t = p_t - [(1-\delta_t)/(1+R_t)] \cdot E_t(p_{t+1}) \qquad \textbf{(1.7)}$$

where uc_t is the user cost of holding a stock of durable for the period t, δ_t is the depreciation rate, R_t is the nominal interest rate and p_t is the price of new durables. We also followed FHS in generating user cost data under two benchmark expectation models: static expectations – the expected price one period ahead is simply today's figure – and perfect foresight[30]. We will return to the robustness of our findings with regard to this assumption when we discuss the test results. To calculate the depreciation rate, we used the ratio of current-value figures on depreciation for each subcategory of durables to its current-value gross stocks at end of periods (net stock plus depreciation). By doing so, we benefited from BEA's recent methodological change described above.

---

[28]Katz and Herman (1997) mentioned such assumption, but made no further comments on how the depreciation profiles have actually changed for a same class of assets, across vintages.

[29]Both SW and FHS follow Diewert (1974) in the adequacy of calculating user costs rather than using prices of new durables for this matter. See Fleissig (1993) for details on the assumptions underlying expression (5).

[30]FHS actually also assumed that consumers would estimate ARIMA models to try and predict next period's prices, but results were pretty much the same as the ones under perfect foresight.

Having described how the depreciation rate is used in our calculation of user costs, the impact of the new method on our analysis becomes somewhat trivial: as it tended to generate larger depreciation rates, all else equal, the new method implied higher user costs of durables – see expression (1.7) above. Figures 1.4 through 1.6 confirm just that, as annual estimates of the user costs for all subcategory of durables are shown to be higher with the new data than with the old series. Therefore, this particular methodological change may affect the conclusions on revealed preference analysis of aggregates for the same reason that the adoption of superlative indices can: as it changes our perception of "prices" for a subset of consumption subcategories, it alters relative prices for the whole set of goods investigated.

Finally, notice that this "user cost approach" explicitly assumes that consumer's expenditures on those services at each period are *proxied* by the product of user costs and current stocks. Alternatively, one could adopt prices and quantities of current purchases of new durables. However, this would imply disregarding the fact that durables purchased currently are expected to provide services for multiple periods. Since the available tools for the revealed preference analysis are essentially built upon a static framework, the approach described before has been standard in the literature. We interpret it as a convenient way to convert dynamic choices concerning stocks of durables into static choices on periodic flows of services that the stocks provide.

1.3.4 The new consumption subcategory

The last methodological change with a possible impact on our analysis is a classificational one; starting with the 1999 Comprehensive Revision of NIPA tables, rather than being a component of the residual series *Other Services*, expenditures on recreation have been treated as a subcategory of the major aggregate of *Services*. Its share on total consumption expenditure increased significantly over the last 40 years, changing from 2% in 1959 to 3.9% in 2000; the boost is partially due to considerable

expansions of casino gambling – ultimately attributed to the increased number of jurisdictions where such activities were legal – and of cable television services (Moran and McCully, 2001).

To understand the potential impact of this classificational change, one must refer back to a central motivation in FHS. As discussed in that paper, the adoption of major consumption aggregates in weak separability tests imposes some *a priori* separability structure involving all subcategories; recall that their tests ended up rejecting the hypothesis of a weakly separable aggregate of all durable goods. The same reasoning can be applied to the treatment of this particular subcategory (recreational services) as a component of the residual "Other services". Recreational services might very well be nonseparable from the allocation of resources (time) to leisure, a hypothesis that simply could not be investigated with data built under the old methodology (we will return to that possibility in section 1.5). However unlikely, this is also a possibility for eventual discrepancies between test results using new and old data, which we discuss next.

1.4 Test implementation and results

1.4.1 GARP-consistency and weak separability of consumption goods only

Our first step is the analysis of consumption aggregates alone – excluding leisure – under NIPA's old and new methodologies, so that we can evaluate how the adoption of revised data and a distinct (lower) susceptibility to substitution bias may affect those tests and lead us to reject/confirm previous findings.

Table 1.2 presents GARP violations and estimated efficiency levels of quarterly and annual consumption data. Under NIPA's old methodology, samples start at 1929 (annual data) and 1959 (annual and quarterly data) and both end at 1990, whereas under the new methodology more recent observations are also available. With quarterly data, our test found no evidence of GARP violations over periods starting in 1959, regardless

of sample size, data methodologies or expectation models adopted. Even though these results confirm in general lines the evidence found in FHS, an interesting discrepancy stands out in the case of perfect foresight expectations; FHS reported violations to GARP in both quarterly and annual data, situated around 1980-1981, for which there seemed to be no apparent economic explanation. We will return to this issue later.

The availability of observations dating back to 1929 is clearly an advantage of annual data in our current evaluation of the impacts of NIPA's methodological changes. It is a well known fact that fixed-weight indices and the real aggregates calculated from them are more susceptible to substitution bias as the investigated periods get more distant from the year-basis of the series (1987=100). As expected, our results under NIPA's old methodology change drastically as annual data allow the inclusion of observations for the 1929-1959 period. We found no violations to GARP over the more recent 1959-1990 period, regardless of the specific expectation mechanism adopted, but up to 12 violations (perfect foresight model) when the earlier observations are included[31]. The highest level of efficiency was associated with the static expectations model (99.85%), resulting from only 2 violations over the 1929-1990 period; it suggests that no inconsistency would have been found if one were willing to relax very little the way revealed preferred bundles are computed. However, this "near-optimizing" behavior does not imply the existence of a (stable) well-behaved utility function rationalizing the data, and the use of the whole sample of data under NIPA's old methodology in estimation exercises is therefore questionable.

As for GARP consistency of figures calculated under the new methodology, only the dataset assuming perfect-foresight calculation of user costs of durables was found inconsistent with the maximization model over the full sample (1929-2000) – despite the much smaller number of violations and higher efficiency level when compared to the same model under NIPA's old methodology. Under static expectations, our test indicates no violations to GARP at all, regardless of the sample range. Finally, we found no

[31]As pointed out by Douglas Pearce, war years may also play an important role on GARP inconsistency of the full sample.

evidence whatsoever of a break-point around 1980-1981 even with annual data, as reported by FHS. We attribute such discrepancy to minor differences in the data used to calculate user costs of durables[32].

In sum, the results in table 1.2 are consistent with the view that the substitution bias can at least partially explain the failure to obtain GARP consistency of consumption expenditure in previous studies. Correcting for substitution bias, as occurred with the adoption of Fisher indices in NIPA tables, does tend to make larger samples of data consistent with the utility maximization model.

Table 1.3 shows the results of weak separability tests for the largest common samples of GARP-consistent data, at each frequency, under both methodologies. The alternative separability structures are organized in groups; the first structure tends to be the most restrictive one, as it assumes mutual weak-separability of two or three groups of consumption subcategories. Take S1, for example: the first proposed separability structure assumes mutual weak-separability of nondurables, durables and services. If just one of these groups does not pass necessary conditions for weak separability, the whole structure will be rejected. On the other hand, if S1 passes both necessary and sufficient conditions tests, results for S1(a) and S1(b) will obviously be uninformative.

In total, we present results for seven separability structures previously investigated in the literature[33], indicating whether such structures pass necessary, sufficient or none of the conditions. Once more, if structures pass tests for none or both conditions, we can make conclusions on the existence of weakly separable preferences rationalizing the data; if they pass only necessary conditions, however, results are inconclusive.

The first (upper) half of table 1.3 shows the weak-separability results of annual aggregates. The results do not seem to differ excessively as data built under old and new methodologies are adopted: there is a large number of separability structures passing tests for both necessary and sufficient conditions, regardless of methodology or

[32]FHS assume a fixed depreciation rate for each durable subcategory throughout the sample, calculated as the reciprocal of numbers of year-life, while we fit rates for each year. See more details in Appendix 1.

expectation mechanism. Particularly interesting, the assumption of nondurables and services as a weakly separable group passes tests for necessary and sufficient conditions, regardless of expectations mechanisms or data construction methodology. However, a striking disparity concerns the structures assuming nondurables as a weakly separable group; confirmed with the data built under the old methodology, it does not pass the sufficiency test with the new data, regardless of the expectation model adopted[34].

As we move to the second part of table 1.3 and study quarterly aggregates, it is clear that most of the proposed separability structures do not pass tests for necessary and sufficient conditions, regardless of data methodology or expectation model. The exceptions occur if one assumes a weakly separable group of nondurables and services, for data built under the new methodology and from a static expectation model. Notice also that all structures assuming weakly separability of durables are conclusively rejected with the old data, which confirms precisely the conclusions of FHS; with the new data, however, such clear rejection only occurs as one assumes a weakly separable group of nondurables. Overall, conclusions on weak separability of consumption goods only (excluding leisure) differ significantly as one investigates datasets built under new and old methodologies.

1.4.2 GARP consistency and weak separability of aggregates, leisure included

We finally examine our complete dataset, using once more NIPA's old and new figures and including leisure/labor choices into the representative consumer's maximizing behavior. Due to the limited availability of labor data, the largest sample of new and old data starts at 1964 and ends at 1990. Both quarterly and annual data passed

[33]We abandoned some of FHS's separability structures, which relied on even more disaggregated data.

[34]Consumption Asset Pricing Models have traditionally adopted both nondurables alone and the sum of nondurables and services in investigations of intertemporal substitution and risk aversion through Euler equation estimations; see Hansen and Singleton (1982), Heaton (1995) and Stock and Wright (2000).

the GARP-consistency test over this more recent period, regardless of methodology or expectation mechanisms adopted in the calculation of user costs.

Table 1.4 presents the results of necessary and sufficient conditions tests for 11 different separability structures. Just like in table 3, the first separability structure of each of the 5 major groups tends to be the most restrictive one, assuming mutual weak separability of many subsets of goods. As a general conclusion, our findings indicate that empirical researchers may actually benefit from a great flexibility in the choice of separability structures in the analysis of recent annual U.S. aggregate data: all proposed structures passed tests for necessary and sufficient conditions, regardless of data methodology or expectation mechanisms. Major aggregates of annual consumption expenditures can be treated either as a single good weakly separable from leisure (SEP3), or as a set of mutually weakly-separable aggregates (SEP1). Assuming nondurables and services as weakly separable from durables and leisure is another common hypothesis in previous empirical studies that found support in our test for annual data.

Inspecting results in the second half of table 1.4, one can see that results with new or old data are markedly different: the use of old data implied no definite rejection, only a couple of structures passing necessary and sufficient condition tests and a majority of inconclusive results. The results of tests performed on new data have shown that conclusions are indeed very sensitive to the choice of expectation mechanisms; many more separability structures passed tests for necessary and sufficient conditions under static expectations than with perfect foresight. Exceptions were SEP1(b) admitting nondurables only, SEP1(c) for services only, and SEP(4), which defines a presumed weakly separable aggregate of nondurables, services and leisure; this last separability structure was the only one passing both tests for necessary and sufficient conditions in all cases, regardless of data methodology or frequency. Such results indicate that

empirical studies of representative agent models using quarterly data should not assume that consumption goods are weakly separable from leisure[35].

The next table (1.5) shows results on weak separability tests for the largest sample of the new data, as a final piece of evidence to guide future choices of separability structures in empirical works. Results with annual data were unchanged with the inclusion of data for the period 1991-2000. However, with quarterly data, only separability structures including a large number of consumption subcategories passed tests for both necessary and sufficient conditions. Once more, the only separability structure passing both tests for necessary and sufficient conditions in all cases, regardless of data methodology or frequency was SEP(4), strongly indicating that nondurables and services should not be treated as separable from leisure. Our general conclusion remains, nevertheless, that empirical researchers may have greater flexibility in choosing separability structures for the study of annual aggregates than in the investigations of quarterly models, where the majority of the tests points to inconclusive results.

1.5 Robustness check and alternative separability assumptions

Before presenting our final conclusions, it is worth extending the current analysis to deal with three potentially relevant aspects of previous empirical studies: (i) the use of Varian's own software, rather than our updated codes; (ii) Hahm's (1998) hypothesis on the weak separability of nondurables and services excluding "Housing services"(S1), which would be inevitably associated with significant (unobservable) adjustment costs; and (iii) the possibility that recreational services (S5) and leisure can form a subgroup weakly separable from other consumption subcategories.

[35]See Mankiw, Rotemberg and Summers (1985), one of the relatively few papers assuming nonseparability of leisure from consumption goods in representative agent models of intertemporal substitution.

Tables 1.3b, 1.4b and 1.5b at the end of this chapter are "mirror" tables, built with results obtained from Varian's NONPAR software for comparison with our own tables 1.3, 1.4 and 1.5. Table 1.3b, for example, has results on the investigation of the same separability structures and datasets considered in table 1.3, the single distinction being the use of older software. Whenever the sample size limitation permitted ($n<75$) we used Varian's codes in their DOS version, which was last revised and updated in 1991; otherwise, we adopted an even older Pascal version, compiled using Prospero's Extended Pascal Compiler. Comparing our tables with the "NONPAR-mirror" ones, there are many results that differ, even though an overall finding remains: annual data are more likely to pass GARP and weak separability tests than series at the quarterly frequency. An example showing how the use of Varian's software may lead to incorrect or imprecise conclusions can be extracted from the very first few entries of table 1.3b. Notice that his software does not reject mutual weak separability of all major consumption aggregates at the annual frequency (1st half of the table 1.3b, structure S1), with the restrictive structure passing both necessary and sufficient conditions; subsequently, when the weak separability of only one of those major categories is tested – nondurables, S1(b) – only the necessary condition is met and the test result is inconclusive. The same kind of inconsistency occurs with quarterly data, as shown in the second half of table 1.4b. SEP1 passes both necessary and sufficient conditions with data built under the old methodology and perfect foresight. Subsequently, less restrictive separability structures fail to obtain sufficient condition (SEP1(a),(b),(c)). Those findings are consistent with Fleissig and Whitney's (2003) claim that a more efficient search for numbers satisfying the Afriat inequalities was needed.

As for the second aspect mentioned above, Hahm (1998) argued that expenditures on housing services should be subtracted from the overall consumption aggregate (including nondurables and services) that is typically adopted in empirical models of intertemporal substitution. The reasoning is simply that renters and homeowners face nontrivial adjustment costs which prevent them from responding optimally to short-run changes in interest rates. The procedure involves assuming that

the remaining categories of nondurables and services actually form a weakly separable group of goods in the representative utility function. To examine this alternative preference structure, we reran our tests with the samples investigated in table 1.5 but excluding housing services from the supposedly weakly separable groups in the structures identified as SEP2 – all of them involving the subcategories of nondurables and services as components of a single aggregate. The test results were unaffected; necessary and sufficient conditions for weak separability were met with annual data, whereas results with quarterly series were either inconclusive or indicated the inexistence of well-behaved subutility functions rationalizing those subsets of data. Therefore, the adoption of Hahm's alternative consumption aggregate in studies using high frequency data can be considered as inadequate as the standard grouping of nondurables and services, in terms of their possible nonseparability from other goods.

Finally, we also studied slightly changed versions of the separability structures discussed in table 1.5, with recreational services and leisure being treated as the arguments of a subutility function. The results – not reported here – tended to reject the existence of weakly separable group with those two goods only. We found inconclusive results more often with annual than with quarterly data, and most of the separability structures did not pass even the necessary condition in the second case, regardless of the expectation model used to calculate user costs of durables.

1.6 Final remarks and conclusions

This first chapter contains new results that contradict previous findings in the literature. The recent methodological changes in data did affect conclusions concerning both GARP-consistency and weak separability of aggregates. Specifically concerning the existence of well-behaved utility functions rationalizing the data, longer samples built under the new methodology showed no GARP-violations; over our largest sample (1929-2000) the overall (Afriat) efficiency level of consumption allocations was clearly higher with the new data than with data built under the old methodology. The fact that those efficiency levels were closer to 1 with the new data (i.e., close to 100% efficiency) indicates that no violations would be detected if one we were willing to admit that consumers can act indifferently to bundles that cost approximately the same. In this sense, the discussed methodological improvements indeed tended to make series of per capita consumption allocations more likely to be consistent with the representative utility maximization model.

As far as we know, this was the first time GARP-consistency and weak separability tests were performed on consumption data generated under NIPA's new methodology. In fact, results were specially sensitive to the adoption of new data as longer samples were considered, which is consistent with the well-known fact that the fixed-weight indices used to generate the old datasets are increasingly susceptible to substitution bias as the investigated periods get further away from the year-basis. Therefore, the substitution bias is a plausible explanation for GARP violations in long annual series of U.S. consumption data. Available only over more recent periods, quarterly series seem to be consistent with GARP regardless of the data construction methodology.

As for weak separability tests, our results with revised data corroborate only part of the assumptions frequently adopted in the empirical literature. Specially concerning the treatment of nondurables and services as weakly separable from other goods, our results tended to confirm such separability structure in the analysis of consumption

goods only, which implicitly imposed separability from leisure. However, as leisure was included in the set of goods, the previously maintained assumption often did not pass the sufficiency test. A weakly separable group of nondurables, services and leisure was found to meet both necessary and sufficient conditions, regardless of data frequency. In any case, NIPA's methodological changes were also found relevant for the analysis of weak separability of goods, regardless of the inclusion of leisure.

The next three chapters will explore research topics that are closely related to our findings here. Chapter 2 investigates whether the new NIPA methodologies also affect previous conclusions about the impacts of temporal aggregation on time series properties of data, as in Rossana and Seater (1995). In chapter 3 we investigate Varian's nonparametric framework itself; we will examine whether and to what extent our general nonrejection of GARP-consistency and weak separability hypotheses within annual datasets can be attributed to a positive correlation between the power of the test and the data frequency. In chapter 4 we improve upon the standard description of leisure choices (adopted in the current chapter) with the explicit consideration of a fourth use of time – besides work, leisure and biological imperatives such as sleeping and eating: the time dedicated to the production of nonmarket household production, which also implies consumption flows unaccounted for in official statistics.

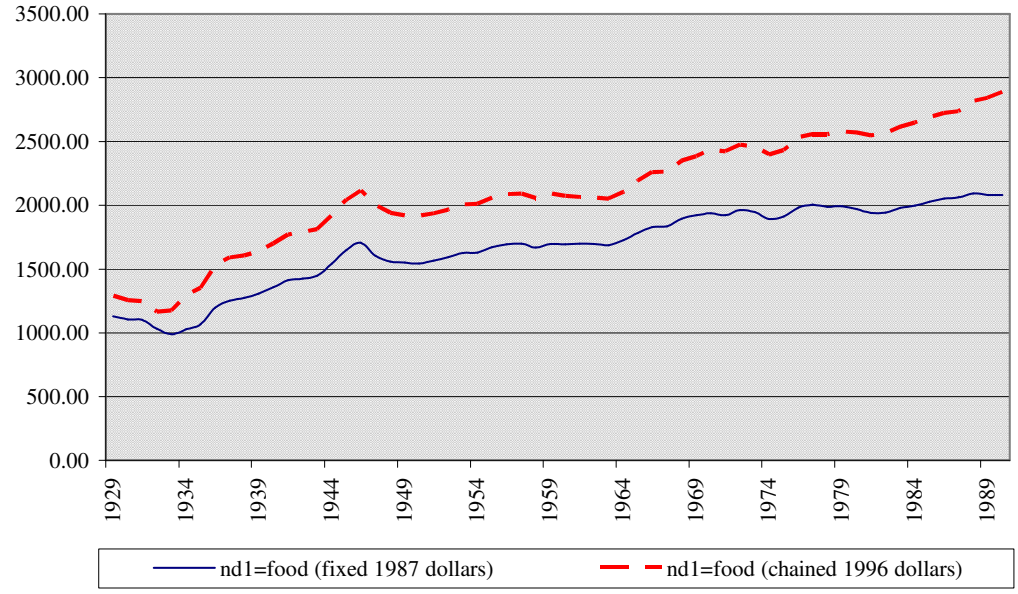

Figure 1.1: NIPA data and the evolution of real quantities under different methodologies

Real per capita expenditure on food, different methodologies (1929-1990)

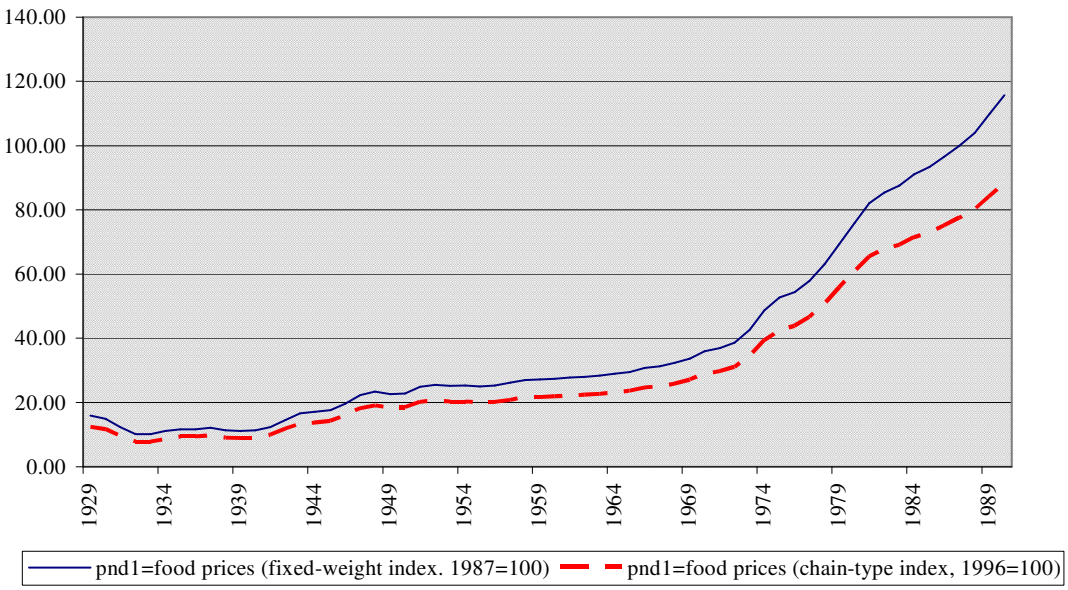

Figure 1.2: NIPA data and the evolution of prices under different methodologies

Indice of food prices, different methodologies (1929-1990)

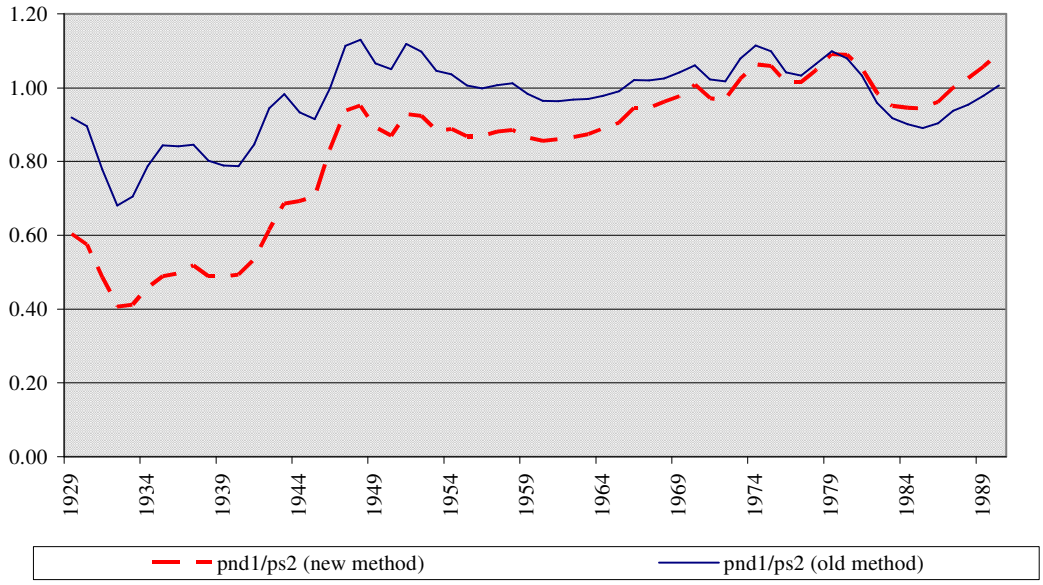

Figure 1.3: Evolution of real relative prices under different methodologies

Food vs. household operations (ratio of price indice, 1929-1990)

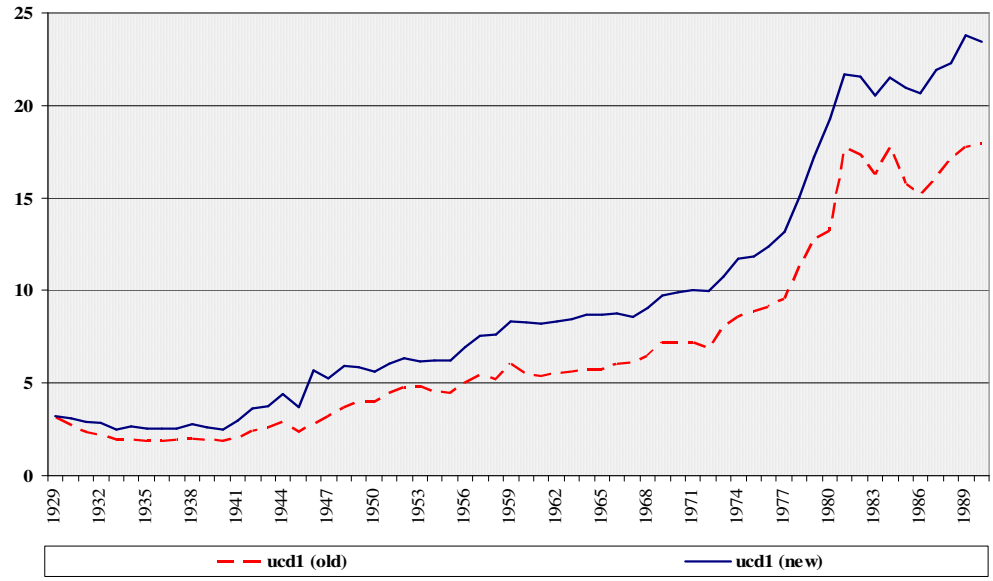

Figure 1.4: Old and new user costs of motor vehicles & parts (D1), static expectations

Values are normalized to 1996 prices of new durables (1996=100)
(1929-1990)

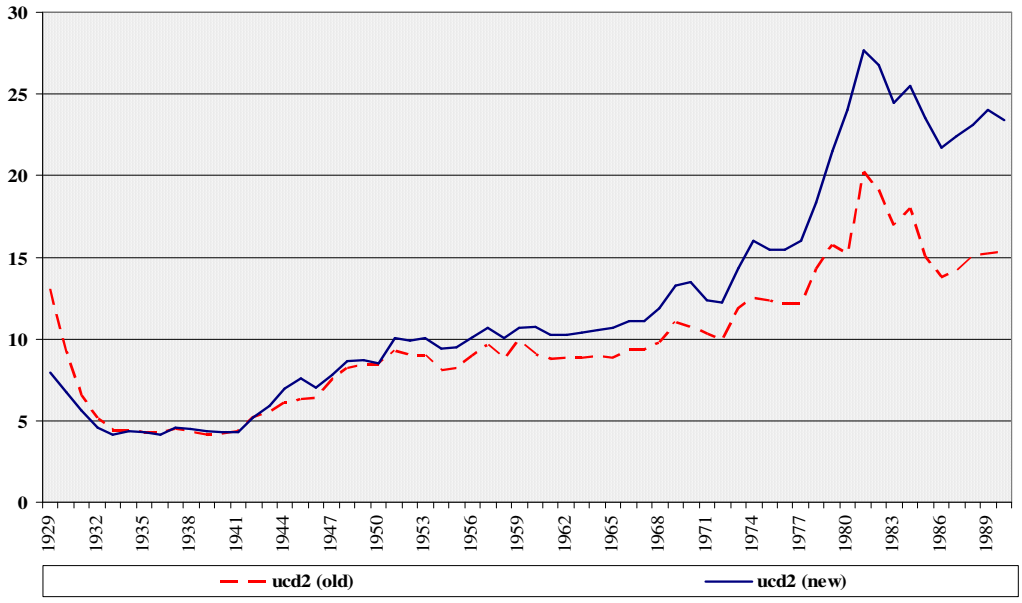

Figure 1.5: Old and new user costs of furniture and household equipments (D2)
Values are normalized to 1996 prices of new durables (1996=100)
(1929-1990)

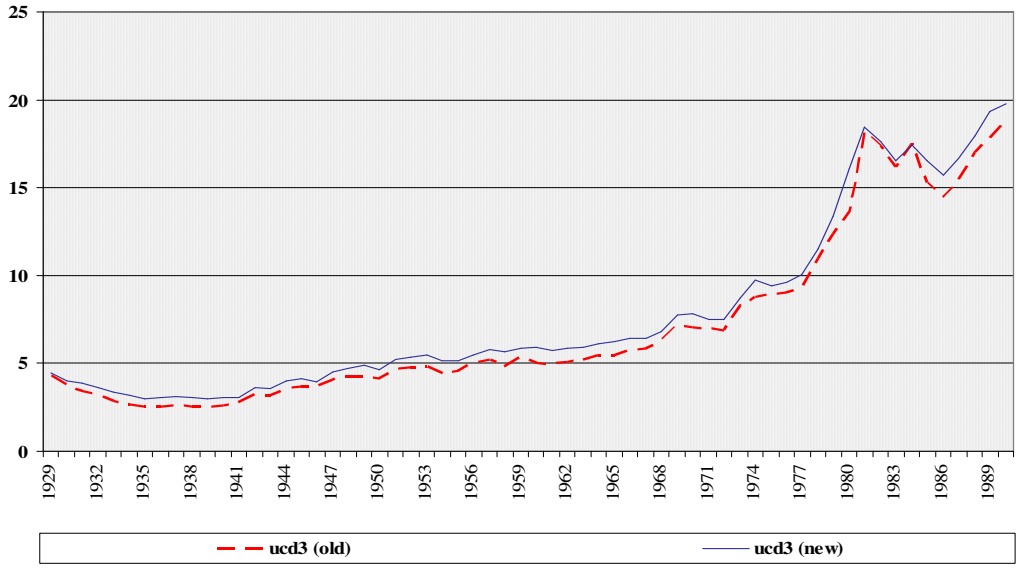

Figure 1.6: Old and new user costs of other durables (D3), static expectations
Values are normalized to 1996 prices of new durables (1996=100)
(1929-1990)

Table 1.1: Depreciation Rates and Profiles under NIPA's Old and New Methodologies
10-year service life (L) hypothetical equipment; 1.65 declining-balance depreciation (dep. rate=1.65/L)

Period	Straight-Line Depreciation Profile		Geometric Pattern Dep. Profile	
	(% of real value)	dep rate (%)	(% of real value)	dep. rate (%)
0	100.00		100.00	
1	90.00	10.00	83.50	16.50
2	80.00	11.11	69.72	16.50
3	70.00	12.50	58.22	16.50
4	60.00	14.29	48.61	16.50
5	50.00	16.67	40.59	16.50
6	40.00	20.00	33.89	16.50
7	30.00	25.00	28.30	16.50
8	20.00	33.33	23.63	16.50
9	10.00	50.00	19.73	16.50
10	0.00	100.00	16.48	16.50

Table 1.2: GARP violations under NIPA's old and new methodologies

	Expectations:	Quarterly Data		Annual Data	
		(1959:I - 1990:IV)	**(1959:I - 2000:IV)**	**(1959-1990)**	**(1929-1990)**
Old Methodology: (Fixed-Weight Indexes)	Stat. Expectations	no violation, e =1.0000	n.a.	no violation, e =1.0000	2 violations, e =0.9985
	Perf. Foresight	no violation, e =1.0000	n.a.	no violation, e =1.0000	12 violations, e =0.9896
New Methodology (Chain-Type Indexes)	Stat. Expectations	no violation, e =1.0000	no violation, e =1.0000	no violation, e =1.0000	no violation, e =1.0000 (*)
	Perf. Foresight	no violation, e =1.0000	no violation, e =1.0000	no violation, e =1.0000	2 violations, e =0.9983 (*)

(*) Exact same results were found for the available larger sample 1929-2000.

Table 1.3: Weak separability of consumption aggregates (excluding leisure)

Annual Data (1947-1990)		Data under NIPA's old methodology		Data under NIPA's new methodology[1]	
		Static Expectations	Perfect Foresight	Static Expectations	Perfect Foresight
Separability Structure2:					
S1	U [u_1(ND) , u_2(S) , u_3(D)]	N&S	N&S	N	N
(a)	U [ND , S , u_1(D)]	N&S	N&S	N&S	N&S
(b)	U [u_1(ND) , S , D]	N&S	N&S	N	N
S2	U [u_1(ND,S) , u_2(D)]	N&S	N&S	N&S	N&S
(a)	U [u_1(ND,S) , D]	N&S	N&S	N&S	N&S
(b)	U [u_1(ND,S) , u_2(D_1) , u_3(D_2,D_3)]	N&S	N&S	N&S	N&S
(c)	U [u_1(ND,S) , u_2(D_1,D_2) , u_3(D_3)]	N&S	N&S	N&S	N&S
S3	U [u_1(ND) , u_2(S,D)]	N&S	N	N	N

Quarterly Data (1959:I-1990:IV)		Data under NIPA's old methodology		Data under NIPA's new methodology[1]	
		Static Expectations	Perfect Foresight	Static Expectations	Perfect Foresight
Separability Structure[2]:					
S1	U [u_1(ND) , u_2(S) , u_3(D)]	X	X	N	X
(a)	U [ND , S , u_1(D)]	X	X	N	X
(b)	U [u_1(ND) , S , D]	N	N	N	N
S2	U [u_1(ND,S) , u_2(D)]	X	X	N	X
(a)	U [u_1(ND,S) , D]	N	N	N&S	N
(b)	U [u_1(ND,S) , u_2(D_1) , u_3(D_2,D_3)]	N	N	N&S	N
(c)	U [u_1(ND,S) , u_2(D_1,D_2) , u_3(D_3)]	N	X	X	X
S3	U [u_1(ND) , u_2(S,D)]	N	N	N	X

Notes: (1) NIPA's new methodology relies on a chain-index (Fisher), rather than more traditional fixed-weight indexes; (2) ND, S and D stand for disaggregated subsets of Nondurables (ND_1,…, ND_5), Services (S_1,…,S_6) and Durables (D_1,..,D_3), respectively. (3) Each separability structure can pass the necessary condition (N), necessary and sufficient conditions (N&S) or none (X) of the conditions for weak separability.

Table 1.4: Weak separability of macroeconomic aggregates (including leisure)

Annual Data (1964 - 1990)		Data under NIPA's old methodology		Data under NIPA's new methodology[1]	
		Static Expectations	Perfect Foresight	Static Expectations	Perfect Foresight
Separability Structure[2]:					
SEP1	U [u_1(ND) , u_2(S) , u_3(D) , u_4(L)]	N&S	N&S	N&S	N&S
(a)	U [u_1(D) , ND , S , L]	N&S	N&S	N&S	N&S
(b)	U [u_1(ND) , S , D , L]	N&S	N&S	N&S	N&S
(c)	U [u_1(S) , ND , D , L]	N&S	N&S	N&S	N&S
(d)	U [u_1(ND) , u_2(S) , u_3(D_1) , u_4(D_2,D_3) , L]	N&S	N&S	N&S	N&S
(e)	U [u_1(ND) , u_2(S) , u_3(D_1,D_2) , u_4(D_3) , L]	N&S	N&S	N&S	N&S
SEP2	U [u_1(ND,S) , u_2(D) , L]	N&S	N&S	N&S	N&S
(a)	U [u_1(ND,S) , D , L]	N&S	N&S	N&S	N&S
(b)	U [u_1(ND,S) , u_2(D_1) , u_3(D_2,D_3) , L]	N&S	N&S	N&S	N&S
(c)	U [u_1(ND,S) , u_2(D_1,D_2) , u_3(D_3) , L]	N&S	N&S	N&S	N&S
SEP3	U [u_1(D,ND,S) , L]	N&S	N&S	N&S	N&S
SEP4	U [u_1(ND,S,L) , D]	N&S	N&S	N&S	N&S
SEP5	U [u_1(D,S,L) , ND]	N&S	N&S	N&S	N&S

Quarterly Data (1964:I - 1990:IV)		Data under NIPA's old methodology		Data under NIPA's new methodology[1]	
		Static Expectations	Perfect Foresight	Static Expectations	Perfect Foresight
Separability Structure[2]:					
SEP1	U [u_1(ND) , u_2(S) , u_3(D) , u_4(L)]	N	N	N	X
(a)	U [u_1(D) , ND , S , L]	N&S	N&S	N&S	X
(b)	U [u_1(ND) , S , D , L]	N	N	N&S	N&S
(c)	U [u_1(S) , ND , D , L]	N	N	N&S	N&S
(d)	U [u_1(ND) , u_2(S) , u_3(D_1) , u_4(D_2,D_3) , L]	N	N	N	N
(e)	U [u_1(ND) , u_2(S) , u_3(D_1,D_2) , u_4(D_3) , L]	N	N	X	X
SEP2	U [u_1(ND,S) , u_2(D) , L]	N	N	N&S	X
(a)	U [u_1(ND,S) , D , L]	N	N	N&S	N
(b)	U [u_1(ND,S) , u_2(D_1) , u_3(D_2,D_3) , L]	N	N	N&S	N
(c)	U [u_1(ND,S) , u_2(D_1,D_2) , u_3(D_3) , L]	N	N	X	X
SEP3	U [u_1(D,ND,S) , L]	N	N	N&S	N
SEP4	U [u_1(ND,S,L) , D]	N&S	N&S	N&S	N&S
SEP5	U [u_1(D,S,L) , ND]	N	N	N	X

Notes: (1)NIPA's new methodology relies on a chain-index (Fisher), rather than more traditional fixed-weight indexes; (2)ND, S, D and L stand for disaggregated sets of Nondurables (ND_1,…, ND_5), Services (S_1,…,S_6), Durables (D_1,..,D_3) and Leisure, respectively. (3) Each separability structure can pass the necessary condition (N), necessary and sufficient (N&S) or none (X) of the conditions for weak separability.

Table 1.5: Weak separability of macroeconomic aggregates, leisure included (1964-2000)

Annual Data (1964 - 2000)

Separability Structure[2]:		Static Expectations	Perfect Foresight
SEP1	U [u_1(ND) , u_2(S) , u_3(D) , u_4(L)]	N&S	N&S
(a)	U [u_1(D) , ND , S , L]	N&S	N&S
(b)	U [u_1(ND) , S , D , L]	N&S	N&S
(c)	U [u_1(S) , ND , D , L]	N&S	N&S
(d)	U [u_1(ND) , u_2(S) , u_3(D_1) , u_4(D_2,D_3) , L]	N&S	N&S
(e)	U [u_1(ND) , u_2(S) , u_3(D_1,D_2) , u_4(D_3) , L]	N&S	N&S
SEP2	U [u_1(ND,S) , u_2(D) , L]	N&S	N&S
(a)	U [u_1(ND,S) , D , L]	N&S	N&S
(b)	U [u_1(ND,S) , u_2(D_1) , u_3(D_2,D_3) , L]	N&S	N&S
(c)	U [u_1(ND,S) , u_2(D_1,D_2) , u_3(D_3) , L]	N&S	N&S
SEP3	U [u_1(D,ND,S) , L]	N&S	N&S
SEP4	U [u_1(ND,S,L) , D]	N&S	N&S
SEP5	U [u_1(D,S,L) , ND]	N&S	N&S

Quarterly Data (1964:I - 2000:IV)

Separability Structure[2]:		Static Expectations	Perfect Foresight
SEP1	U [u_1(ND) , u_2(S) , u_3(D) , u_4(L)]	X	X
(a)	U [u_1(D) , ND , S , L]	N	X
(b)	U [u_1(ND) , S , D , L]	X	X
(c)	U [u_1(S) , ND , D , L]	N	N
(d)	U [u_1(ND) , u_2(S) , u_3(D_1) , u_4(D_2,D_3) , L]	X	X
(e)	U [u_1(ND) , u_2(S) , u_3(D_1,D_2) , u_4(D_3) , L]	X	X
SEP2	U [u_1(ND,S) , u_2(D) , L]	N	X
(a)	U [u_1(ND,S) , D , L]	N	N
(b)	U [u_1(ND,S) , u_2(D_1) , u_3(D_2,D_3) , L]	N	N
(c)	U [u_1(ND,S) , u_2(D_1,D_2) , u_3(D_3) , L]	X	X
SEP3	U [u_1(D,ND,S) , L]	N&S	N
SEP4	U [u_1(ND,S,L) , D]	N&S	N&S
SEP5	U [u_1(D,S,L) , ND]	N	N

Notes: (1)ND, S, D and L stand for disaggregated sets of Nondurables (ND1,…, ND5), Services (S1,…,S6), Durables (D_1,..,D_3) and Leisure, respectively; (2) Each sep. structure can pass the necessary condition (N), necessary and sufficient (N&S) or none (X) of the conditions for weak separability.

Table 1.3b: Weak separability of consumption aggregates (excluding leisure) - NONPAR

Annual Data (1947-1990)		Data under NIPA's old methodology		Data under NIPA's new methodology[1]	
Separability Structure[2]:		Static Expectations	Perfect Foresight	Static Expectations	Perfect Foresight
S1	U [u_1(ND) , u_2(S) , u_3(D)]	N&S	N&S	N&S	N
(a)	U [ND , S , u_1(D)]	N&S	N&S	N&S	N&S
(b)	U [u_1(ND) , S , D]	N	N	N	N
S2	U [u_1(ND,S) , u_2(D)]	N&S	N&S	N&S	N&S
(a)	U [u_1(ND,S) , D]	N&S	N&S	N&S	N&S
(b)	U [u_1(ND,S) , u_2(D_1) , u_3(D_2,D_3)]	N&S	N&S	N	N&S
S3	U [u_1(ND) , u_2(S,D)]	N&S	N&S	N	N

Quarterly Data (1959:I-1990:IV)		Data under NIPA's old methodology		Data under NIPA's new methodology[1]	
Separability Structure[2]:		Static Expectations	Perfect Foresight	Static Expectations	Perfect Foresight
S1	U [u_1(ND) , u_2(S) , u_3(D)]	X	X	N	X
(a)	U [ND , S , u_1(D)]	X	X	N	X
(b)	U [u_1(ND) , S , D]	N	N	N	N
S2	U [u_1(ND,S) , u_2(D)]	X	X	N	X
(a)	U [u_1(ND,S) , D]	N	N	N	N
(b)	U [u_1(ND,S) , u_2(D_1) , u_3(D_2,D_3)]	N	N	N&S	N
S3	U [u_1(ND) , u_2(S,D)]	N&S	N	N	X

Notes: (1) NIPA's new methodology relies on a chain-index (Fisher), rather than more traditional fixed-weight indexes; (2) ND, S and D stand for disaggregated subsets of Nondurables (ND_1,…, ND_5), Services (S_1,…,S_6) and Durables (D_1,..,D_3), respectively. (3) Each separability structure can pass the necessary condition (N), necessary and sufficient conditions (N&S) or none (X) of the conditions for weak separability.

Table 1.4b: Weak separability of macroeconomic aggregates (including leisure) - NONPAR

Annual Data (1964 - 1990)		Data under NIPA's old methodology		Data under NIPA's new methodology[1]	
		Static Expectations	Perfect Foresight	Static Expectations	Perfect Foresight
Separability Structure[2]:					
SEP1	U [u_1(ND) , u_2(S) , u_3(D) , u_4(L)]	N&S	N&S	N&S	N&S
(a)	U [u_1(D) , ND , S , L]	N&S	N&S	N&S	N&S
(b)	U [u_1(ND) , S , D , L]	N	N	N&S	N&S
(c)	U [u_1(S) , ND , D , L]	N&S	N&S	N&S	N&S
(d)	U [u_1(ND) , u_2(S) , u_3(D_1) , u_4(D_2,D_3) , L]	N&S	N&S	N&S	N&S
SEP2	U [u_1(ND,S) , u_2(D) , L]	N&S	N&S	N&S	N&S
(a)	U [u_1(ND,S) , D , L]	N&S	N&S	N&S	N&S
(b)	U [u_1(ND,S) , u_2(D_1) , u_3(D_2,D_3) , L]	N&S	N&S	N&S	N&S
SEP3	U [u_1(D,ND,S) , L]	N&S	N&S	N&S	N&S
SEP4	U [u_1(ND,S,L) , D]	N&S	N&S	N&S	N&S
SEP5	U [u_1(D,S,L) , ND]	N&S	N&S	N&S	N&S

Quarterly Data (1964:I - 1990:IV)		Data under NIPA's old methodology		Data under NIPA's new methodology[1]	
		Static Expectations	Perfect Foresight	Static Expectations	Perfect Foresight
Separability Structure[2]:					
SEP1	U [u_1(ND) , u_2(S) , u_3(D) , u_4(L)]	N	N&S	N&S	X
(a)	U [u_1(D) , ND , S , L]	N	N	N	X
(b)	U [u_1(ND) , S , D , L]	N	N	N&S	N&S
(c)	U [u_1(S) , ND , D , L]	N	N	N	N
(d)	U [u_1(ND) , u_2(S) , u_3(D_1) , u_4(D_2,D_3) , L]	N	N&S	N&S	N
SEP2	U [u_1(ND,S) , u_2(D) , L]	N	N	N&S	X
(a)	U [u_1(ND,S) , D , L]	N	N	N	N
(b)	U [u_1(ND,S) , u_2(D_1) , u_3(D_2,D_3) , L]	N	N	N	N&S
SEP3	U [u_1(D,ND,S) , L]	N	N	N	N
SEP4	U [u_1(ND,S,L) , D]	N	N	N	N
SEP5	U [u_1(D,S,L) , ND]	N	N	N	X

Notes: (1)NIPA's new methodology relies on a chain-index (Fisher), rather than more traditional fixed-weight indexes; (2)ND, S, D and L stand for disaggregated sets of Nondurables (ND_1,…, ND_5), Services (S_1,…,S_6), Durables (D_1,..,D_3) and Leisure, respectively. (3) Each separability structure can pass the necessary condition (N), necessary and sufficient (N&S) or none (X) of the conditions for weak separability.

Table 1.5b: Weak separability of macroeconomic aggregates, leisure included (1964-2000), NONPAR

Annual Data (1964 - 2000)

Separability Structure[2]:		Static Expectations	Perfect Foresight
SEP1	$U [u_1(ND) , u_2(S) , u_3(D) , u_4(L)]$	N&S	N&S
(a)	$U [u_1(D) , ND , S , L]$	N&S	N&S
(b)	$U [u_1(ND) , S , D , L]$	N&S	N&S
(c)	$U [u_1(S) , ND , D , L]$	N&S	N&S
(d)	$U [u_1(ND) , u_2(S) , u_3(D_1) , u_4(D_2,D_3) , L]$	N&S	N&S
SEP2	$U [u_1(ND,S) , u_2(D) , L]$	N&S	N&S
(a)	$U [u_1(ND,S) , D , L]$	N&S	N&S
(b)	$U [u_1(ND,S) , u_2(D_1) , u_3(D_2,D_3) , L]$	N&S	N&S
SEP3	$U [u_1(D,ND,S) , L]$	N&S	N&S
SEP4	$U [u_1(ND,S,L) , D]$	N&S	N&S
SEP5	$U [u_1(D,S,L) , ND]$	N&S	N&S

Quarterly Data (1964:I - 2000:IV)

Separability Structure[2]:		Static Expectations	Perfect Foresight
SEP1	$U [u_1(ND) , u_2(S) , u_3(D) , u_4(L)]$	X	X
(a)	$U [u_1(D) , ND , S , L]$	N	X
(b)	$U [u_1(ND) , S , D , L]$	X	X
(c)	$U [u_1(S) , ND , D , L]$	N	N
(d)	$U [u_1(ND) , u_2(S) , u_3(D_1) , u_4(D_2,D_3) , L]$	X	X
SEP2	$U [u_1(ND,S) , u_2(D) , L]$	N	X
(a)	$U [u_1(ND,S) , D , L]$	N	N
(b)	$U [u_1(ND,S) , u_2(D_1) , u_3(D_2,D_3) , L]$	N	N
SEP3	$U [u_1(D,ND,S) , L]$	N&S	N
SEP4	$U [u_1(ND,S,L) , D]$	N&S	N
SEP5	$U [u_1(D,S,L) , ND]$	N	X

Notes: (1)ND, S, D and L stand for disaggregated sets of Nondurables (ND1,…, ND5), Services (S1,…,S6), Durables (D_1,..,D_3) and Leisure, respectively; (2) Each sep. structure can pass the necessary condition (N), necessary and sufficient (N&S) or none (X) of the conditions for weak separability.

Chapter 2

Time series analysis of the new NIPA data: reassessing Rossana & Seater's (1995) findings on the impacts of temporal aggregation

2.1 Motivation

In chapter 1 we showed, among other things, that NIPA's new methodologies ended up correcting the evolution of relative prices for many consumption subcategories. As a result of that and other methodological changes, previous findings concerning the revealed preference analysis of per capita expenditures were changed. We observed longer periods of GARP-consistent data with the new series than with the old ones and we also noticed that the results concerning weak separability of subsets of quarterly data were mostly changed. Now it seems appropriate to consider whether the most important methodological change discussed in that chapter has also affected either the univariate time-series properties of aggregates or previous conclusions about the impacts of temporal aggregation on those properties. We will reapply Rossana and Seater's (1995) method on series originally investigated in that paper and also on prices and real expenditures of consumption expenditures subcategories[36], using data at different frequencies calculated under old and new methodologies.

[36]Besides NIPA data, their study included variables such as unemployment rates, nominal monetary stocks (M2) and interest rates; those series are not reconsidered here simply because we have no knowledge of significant methodological changes in the way those numbers are built.

Our focus now is on the time series analysis of those numbers and, consequently, we must emphasize that the adoption of the new index implies working with a more complex combination of stochastic processes. Consider once more the formulas involved in the estimation of real expenditures under NIPA's old and new methodologies, reproduced here for convenience:

$$q^{LASP} = \sum_{n=1}^{N} P_2^n Q_2^n \times \begin{bmatrix} \left(\sum_{n=1}^{N} P_2^n Q_1^n / \sum_{n=1}^{N} P_2^n Q_2^n \right) \\ \left(\sum_{n=1}^{N} P_2^n Q_2^n / \sum_{n=1}^{N} P_2^n Q_2^n \right) \\ \left(\sum_{n=1}^{N} P_2^n Q_3^n / \sum_{n=1}^{N} P_2^n Q_2^n \right) \end{bmatrix} \quad \textbf{(2.1)}$$

$$q^{FISH} = \sum_{n=1}^{N} P_2^n Q_2^n \times \begin{bmatrix} \left(\left(\sum_{n=1}^{N} P_2^n Q_1^n / \sum_{n=1}^{N} P_2^n Q_2^n \right) \times \left(\sum_{n=1}^{N} P_1^n Q_1^n / \sum_{n=1}^{N} P_1^n Q_2^n \right) \right)^{0.5} \\ \left(\left(\sum_{n=1}^{N} P_2^n Q_2^n / \sum_{n=1}^{N} P_2^n Q_2^n \right) \times \left(\sum_{n=1}^{N} P_2^n Q_2^n / \sum_{n=1}^{N} P_2^n Q_2^n \right) \right)^{0.5} \\ \left(\left(\sum_{n=1}^{N} P_2^n Q_3^n / \sum_{n=1}^{N} P_2^n Q_2^n \right) \times \left(\sum_{n=1}^{N} P_3^n Q_3^n / \sum_{n=1}^{N} P_3^n Q_2^n \right) \right)^{0.5} \end{bmatrix} \quad \textbf{(2.2)}$$

As discussed before, the Fisher-Ideal index is calculated as the geometric average of the traditional Laspeyres index and the less commonly used Paasche index. The second term involves "moving weights" for changes between the current period and some base-year (period 2, in those cases), which makes the adoption of the new index more costly and computationally burdensome, despite its advantages[37]. To our knowledge, there is no theory on how geometric average of possibly nonstationary processes should behave in finite samples – as there is for linear relationships among

[37]Most notoriously, the elimination of the substitution bias and the fact that past observations of price or quantity changes are not altered with the periodical moving of base-years.

multiple unit-root processes, for example[38]. Therefore, it seems appropriate and of interest at this time to answer three empirical questions: (i) whether univariate models fitting the new NIPA series contain longer or more complex lag-structures than data calculated under the old methodology; (ii) whether they are less susceptible to temporary shocks (outliers) that could affect our conclusions about their stationarity, due to an eventual smoothing effect resulting from the averaging of different indices in the new data; and (iii) whether previous findings concerning the impacts of temporal aggregation on time-series properties of data have changed with the introduction of the new methodology. The relevance of this last question will be evident next, as we present Rossana and Seater's (1995) contribution as a starting point to our reassessment of previous findings/conclusions.

2.2 Brief summary of Rossana and Seater's method and conclusions

Rossana and Seater proposed a rather intuitive procedure to evaluate time-series properties of economic variables. The first step was to investigate the presence of unit root processes in all series[39]. They performed the Augmented Dickey Fuller (ADF) unit root tests with and without the inclusion of intercept and trend terms in the test regression. As all series were found integrated of order 1, they fit alternative ARIMA models to each nonstationary series, finally selecting the model specifications with the best performances at each frequency. The list of the model specification considered in that paper is presented below:

[38]We refer to the recent literature on (vector) error-correction models. Hamilton (1994) reports that the general idea was implicit in models advocated by Davidson et al. (1978), which observed that even though consumption and income exhibit unit roots, the linear combination of those variables seemed stationary.

[39]They actually started investigating the presence of seasonal unit roots first, finding none. We have not performed those tests here simply because all series considered in this chapter are seasonally adjusted.

RW : $\triangle x_t = \mu + e_t$ **(2.3)**

IMA(1,1) : $\triangle x_t = \mu + (1 - \beta L).e_t$ **(2.4)**

ARI(1,1) : $(1 - \alpha L).\triangle x_t = \mu + e_t$ **(2.5)**

ARIMA([s] , 1 , [s]) : $(1 - \alpha L^s).\triangle x_t = \mu + (1 - \beta L^s).e_t$ **(2.6)**

ARIMA([1,s] , 1 , [1,s]) : $(1 - \alpha_1 L - \alpha_2 L^s).\triangle x_t = \mu + (1 - \beta_1 L - \beta_2 L^s).e_t$ **(2.7)**

ARI(p,1): $(1 - \sum_{i=1}^{p} \alpha_i L^i).\triangle x_t = \mu + e_t$ **(2.8)**

where p and s represent lags 4 and 12 with quarterly data and 1 and 4 with annual data. Following the same notation in that paper, ARIMA([1,s], 1 , q) means that the series is differenced once, the model includes two autoregressive terms – at lags 1 and s – and from 1 to q moving-average terms. On the first round of regressions they restrict their attention to the specifications above for tractability and because they observed that estimations of the most general specifications ARIMA(p,1,q) often did not converge. Other specifications were tried when none of the models above generated white noises.

The selection method involved choosing the model with the lowest Schwartz information criterion value among all alternative specifications that generated white noise residuals – according to Box-Pierce Q statistics[40]. The goal was to find the simplest model specification fitting data well enough to generate uncorrelated residuals.

They concluded that the loss of information resulting from time aggregation was significant, particularly as one moved from monthly or quarterly to annual data. The elected ARIMA model specifications for data at those frequencies generally had longer (and more complex) lag structures than the ones fitting annual series. The result was considered consistent with what is referred to as Tiao effect: as the level of temporal aggregation increases, a reduced number of autoregressive and moving average

[40]The authors do not report for which lag the Q statistics was calculated; in our exercises we assume that the residuals of a regression are white-noise if such hypothesis cannot be rejected at the 10% significance level considering the first 12 lags. Also worth mentioning, all tests and regressions were run on logged series.

coefficients tends to remain statistically significant. In the limit, the estimated processes behave as IMA(d,d), where d is the order of integration of original series (before the level of temporal aggregation is changed). A second effect discussed in their brief review of the literature on temporal aggregation, the sample size effect, also seemed to be relevant in a few cases; as the sample size is reduced to one fourth with the aggregation of quarterly data into annual series, for example, the estimated standard errors tend to be larger and the observed significance of the moving average terms discussed above would tend to be smaller. As a result, researchers may end up electing the random walk model as the best specification[41].

Among the implications further explored with complementary exercises, the most important one concerns Nelson and Plosser's (1982) conclusion that variations in annual macroeconomic data are dominated by permanent shocks and, consequently, are not of cyclical nature. Rossana and Seater (1995) argued that economic series actually contained far more cyclical variation than one could observe with the analysis of data solely at the annual frequency. The use of quarterly data, instead, was advocated as optimal because the series seemed to preserve most of the dynamic behavior of higher frequency data, without suffering from measurement problems as the ones discussed in Wilcox (1992).

2.3 Best-fitting models and test results

The first finding in this reassessment of Rossana and Seater's conclusions does not need to be reported in a table. With a couple of exceptions that will be discussed later, the presence of a (nonseasonal) unit root was not rejected at the usual significance levels (5 or 1%), regardless of data frequency or the inclusion of constant and/or trend

[41]A third effect predicted by the theory and also reviewed by Rossana and Seater did not find support in their study; known as Brewer effect, it predicts that the order of the autoregressive process (p) tends to

terms in the ADF test regression[42]. Therefore, one of the questions raised in the first section of this chapter is already answered: the series calculated with the new index do not seem to be any less subject than the old ones to temporary shocks that could affect conclusions about their nonstationarity.

We now proceed to report the best-fitting models in multiple datasets, which will permit answering the other two questions, i.e., whether the simple adoption of new indices changes the length/complexity of lag-structures of ARIMA models fitting NIPA series and whether previous findings concerning the impacts of temporal aggregation on time-series properties of data have changed with the new methodology. The tables showing our results contain an additional piece of information that was not reported in the original paper: we also calculated, whenever applicable, the F-statistic and the level of significance [labeled Prob(F-statistics)] at which one could reject the hypothesis that all coefficients in a particular regression – excluding the constant – were null.

2.3.1 Comparing best-fitting models with old and new NIPA series

Table 2.1 reports the best-fitting models for 5 annual series on prices of nondurables over the sample 1929-1990. The choice for those consumption subcategories is due to the fact that the other two major categories of consumption expenditures (durables and services) were subject to multiple methodological changes, rather than only the adoption of a different index in the construction of aggregates[43]. As

---

remain fixed as the data frequency is lowered – by summation of observations – whereas the order of the moving average process (originally, q) approaches p+d+1 in more (time)aggregated series.

[42]We also performed unit root tests on series both before and after applying logs. We observed that the statistical significance of the trend term in the ADF test regressions was sensitive to such procedure. Overall, the hypothesis of nullity of the trend term was rejected often. Nevertheless, in order to generate results comparable to Rossana and Seater's, we restricted attention in this chapter to model specification without deterministic trends that were able to generate white-noise residuals.

[43]Recall from chapter 1 that estimates for the user cost of durables would also be affected by new methodologies in the calculation of depreciation rates, whereas the major category of services is currently broken into 6 subcategories – rather than 5 as in the old data.

the idea is to isolate the impact of that particular methodological change, the choice for nondurables seems the most appropriate one.

Notice initially that in just one case the best-fitting model for series built with different indices coincided, the consumption of fuel oil and coal (ND4). For all other subcategories the elected specifications within the old series had either longer lag structures (ND1,ND2) and/or slightly more complex ones (ND3,ND5) than with the new NIPA data. As striking as this observation may seem, one must be careful to avoid overstating its implications. The superiority of alternative model specifications was marginal most times; the Schwartz information criterion often differed only at the second decimal place for competing models generating white-noise residuals. Additionally, results not reported here indicate that with quarterly data the elected model specification coincided within four of the five series of nondurables subcategories[44].

2.3.2 Best-fitting models for new NIPA data at different frequencies

We now consider the impact of time aggregation on the estimated time-series properties of both price and quantity series of nondurable goods, using the new data only. As throughout this dissertation, we will use per capita rather than aggregate consumption data to study the behavior of real expenditures at multiple frequencies.

Table 2.2 shows the best-fitting specifications for those series over the period 1959:1–2000:4. In general, the elected models for annual series contained either the same or (most often) simpler lag structures than the ones fitting quarterly data. The only exception was the residual subcategory of other nondurable goods (ND5), with more complex lag-structures being found with both price and quantity series at the annual frequency. We could not reject the hypothesis that residuals from the IMA(1,1) model

[44]We chose to report results for annual data because, as discussed in chapter 1, the occurrence of substitution bias in fixed-weight estimates is more likely as observations are further distant from the base-year. Samples of annual data built with both methods are available starting in 1929, whereas quarterly series are only available for postwar periods.

specifications were white-noise processes in all five series of real expenditures, even though those models ended up being elected only twice. This last piece of evidence confirms Rossana and Seater's conclusion about the consistency of the Tiao effect with the overall simpler dynamic behavior of postwar annual data, relatively to the ones fitting quarterly and monthly series.

2.3.3 Reconsidering NIPA series also studied in Rossana and Seater (1995)

We finally replicate Rossana and Seater's analysis using the new NIPA series for the same sample periods and attempt to define/calculate series as close as possible to theirs. Rather than adopting U.S total population to calculate per capita figures as we do in the current research project, Rossana and Seater considered the noninstitutional population including armed forces overseas – data also from the U.S. Census Bureau.

There is, however, one aspect of their series on consumption aggregates that we were not able to reproduce. Concerned with a somewhat neglected durability of a particular subcategory of nondurables, the authors opted for subtracting clothing and shoes (ND2) from the simple sum of real consumption expenditures on nondurable and services, subsequently adding those numbers to the major category of durable goods. The problem in this case is that, under the new methodologies, real expenditures on subcategories do not add up to the corresponding figures for a major consumption category – as discussed previously in chapter 1. One could still start from (real) chained dollar expenditures on all relevant subcategories, excluding ND2, and construct estimates for that alternative aggregate that would be consistent with the new aggregation methodology. However, any discrepancy between our findings and Rossana and Seater's could be attributed to the very adoption of a different aggregation method[45]. Rather than simply disregarding 4 of the 6 NIPA series investigated in the original paper,

[45]See Fleissig, Gallant and Seater (2000) for discussion and evidence on the adoption of different aggregation methods and its impacts on empirical investigations.

we replaced them with the series on the standard aggregates of nondurables and durables, over the same sample period

For convenience, we reproduced in table 2.3 their best-fitting model specifications for all economic variables that we now reconsider. They are: Real GNP, the GNP implicit deflator, nondurables plus services (aggregate & per capita figures) and durables (aggregate & per capita data).

Overall, our results confirm Rossana and Seater's conclusion that temporal aggregation results in significant losses of information about the processes driving the variables at higher frequencies. With the new series, the random walk specification was quite often the best-fitting one for annual data, which does not contrast at all with Rossana and Seater's observation that such specification would also generate white-noise residuals in 5 of the 6 series. As for the evolution of real GNP, the use of the new NIPA figures led to the election of a slightly more complex ARIMA model fitting the data at the quarterly frequency, but not with annual figures; the random walk specification was chosen the best-fitting model for the annual series regardless of data methodology, suggesting the dominance of the sample-size effect in that particular case.

As for the series on consumption expenditures, we did not find model specifications that are hard to conciliate with the ones in the original paper, despite the differences concerning aggregation methods and the very definition of consumption categories. It is worth mentioning in this case, though, that the ADF test did reject the presence of a unit root at 5% significance level under certain circumstances (including constant and trend terms in the test regression, annual series in logs). Since multiple unit-root model specifications still fitted those series reasonably well to generate white-noise residuals, we interpret those results as an artifact of the limited sample size. Note that such interpretation is consistent with the robust nonrejection of the unit-root hypothesis for all subcategories of nondurables within larger samples, as reported in table 2.2.

2.4 Summary of findings

Univariate time series models fitting the new NIPA data did not systematically reveal longer or more complex lag-structures than data calculated under the old methodology. The series calculated from superlative indices did not seem to be any less susceptible to shocks that could affect conclusions about their stationarity than the ones derived from the fixed-price methodology. As for the impact of temporal aggregation on time-series properties of data, Rossana and Seater's conclusion was maintained, specially concerning the consistency of findings with both Tiao and sample size effects and the consequent conclusion about a severe loss of information resulting from time aggregation.

Table 2.1 - Model-fitting results[(1)] for the prices of consumption subcategories using NIPA's old and new annual data (1929-1990)

Variables:	Old method (fixed-weight indices)	New method (chain-type indices)
Food (ND1)	ARIMA(4,1,4) F-statistic: 8.4730 Prob(F-statistic): <0.0001	ARIMA(1,1,1) F-statistic: 18.1428 Prob(F-statistic): <0.0001
Clothing and Shoes (ND2)	ARIMA(4,1,4) F-statistic: 7.9612 Prob(F-statistic): <0.0001	ARI(1,1) F-statistic: 12.8593 Prob(F-statistic): 0.0013
Gasoline and oil (ND3)	ARIMA([1,4] , 1 , [1,4])[†,††] F-statistic: 5.9348 Prob(F-statistic): 0.0005	IMA(1,1)[††] F-statistic: 5.7788 Prob(F-statistic): 0.0283
Fuel oil and coal (ND4)	IMA(1,1) F-statistic: 19.8070 Prob(F-statistic): <0.0001	IMA(1,1) F-statistic: 19.6935 Prob(F-statistic): 0.0041
Other nondurable goods (ND5)	ARIMA(1,1) F-statistic: 22.3580 Prob(F-statistic): <0.0001	IMA(1,1,1) F-statistic: 41.3781 Prob(F-statistic): <0.0001

Notes: (1) Same notation and procedures used in the two previous tables of the current chapter.

Table 2.2 - Model-fitting results[(1)] for consumption subcategories at multiple frequencies (1959-2000)

Variables[(2)]:	Real expenditures		Prices	
	Quarterly data	**Annual data**	**Quarterly data**	**Annual data**
Food (ND1)	ARIMA(1,1,1)†,††	RW†	ARIMA(3,1,3)*	ARIMA(1,1,1)
	F-statistic: 6.5965		F-statistic: 33.5122	F-statistic: 29.2153
	Prob(F-statistic): 0.0018		Prob(F-statistic): <0.0001	Prob(F-statistic): <0.0001
Clothing and Shoes (ND2)	ARIMA([4] , 1 , [4])†,††	ARIMA([4] , 1 , [4])†,††	ARIMA(2,1,2)	ARI(1,1)
	F-statistic: 8.7076	F-statistic: 9.9146	F-statistic: 19.7798	F-statistic: 50.7407
	Prob(F-statistic): 0.0003	Prob(F-statistic): 0.0004	Prob(F-statistic): <0.0001	Prob(F-statistic): <0.0001
Gasoline and oil (ND3)	ARIMA([4] , 1 , [4])†,††	ARIMA([4] , 1 , [4])†,††	ARIMA(3,1,3)	IMA(1,1)††
	F-statistic: 7.9459	F-statistic: 11.4279	F-statistic: 9.8256	F-statistic: 7.7461
	Prob(F-statistic): 0.0005	Prob(F-statistic): 0.0002	Prob(F-statistic): <0.0001	Prob(F-statistic): 0.0083
Fuel oil and coal (ND4)	ARIMA(2,1,2)	IMA(1,1)	ARIMA(1,1,1)	IMA(1,1)††
	F-statistic: 4.2348	F-statistic: 11.2390	F-statistic: 18.2617	F-statistic: 10.7212
	Prob(F-statistic): 0.0027	Prob(F-statistic): 0.0018	Prob(F-statistic): <0.0001	Prob(F-statistic): 0.0022
Other nondurable goods (ND5)	ARI(1,1)†	ARIMA([4] , 1 , [4])†,††	ARI(1,1)	ARIMA(3,1,3)
	F-statistic: 26.8515	F-statistic: 8.8850	F-statistic: 401.2128	F-statistic: 15.3870
	Prob(F-statistic): <0.0001	Prob(F-statistic): 0.0008	Prob(F-statistic): <0.0001	Prob(F-statistic): <0.0001

Notes: (1) Following Rossana and Seater's (1995) notation, we marked with "†" the cases in which the IMA(1,1) specification also generated white-noise residuals, but some other model was preferable according to the Schwartz criterion; similarly, we used "††" to indicate that the Random Walk (RW) specification was acceptable but not preferable; (2) Augmented Dickey-Fuller tests under two specifications - with a constant or with constant and trend terms - were applied to each series; the tests did not reject the presence of a unit root at the 5% significance level, in all cases.

Table 2.3 - Model-fitting results[(1)]: the impacts of new NIPA methodologies

Variables:	Rossana and Seater (1995)		New NIPA data	
	Quarterly data	**Annual data**	**Quarterly data**	**Annual data**
Real GNP (1948-1987)	ARI(1,1)†	RW†	ARIMA([1,4] , 1 , [1,4])† F-statistic: 11.0723 Prob(F-statistic): <0.0001	RW†
GNP Imp. Deflator (1948-1987)	ARIMA(1,1,1)	ARI(1,1)	ARIMA([1,4] , 1 , [1,4]) F-statistic: 35.4592 Prob(F-statistic): <0.0001	ARIMA([1,4] , 1 , [1,4]) F-statistic: 42.8874 Prob(F-statistic): <0.0001
Nondurables+Services (1959-1988)	RW†	IMA(1,1)††	-	-
Per Capita Nondurables+Services (1959-1988)	ARIMA(1,1,1)	IMA(1,1)††	-	-
Durables (1959-1988)	RW†	RW†	-	-
Per Capita Durables (1959-1988)	RW†	RW†	-	-
Addendum:				
Nondurables (chained 1996 $) (1959-1988)	-	-	ARI(1,1)† F-statistic: 10.1798 Prob(F-statistic): 0.0018	RW†,(2)
Per Capita Nondurables (chained 1996 $) (1959-1988)	-	-	ARI(1,1)† F-statistic: 9.3375 Prob(F-statistic): 0.0028	RW†
Durables (chained 1996 $) (1959-1988)	-	-	RW†	RW†,(2)
Per Capita Durables (chained 1996 $) (1959-1988)	-	-	RW†	RW†
Real GNP (1948-2000)	-	-	ARIMA([1,4] , 1 , [1,4]) F-statistic: 12.9747 Prob(F-statistic): <0.0001	RW†,(3)

Notes: (1) Following Rossana and Seater's (1995) notation, we marked with "†" the cases in which the IMA(1,1) specification also generated white-noise residuals, but some other model was preferable according to the Schwartz criterion; similarly, we used "††" to indicate that the Random Walk (RW) specification was acceptable but not preferable; (2) The ADF test rejected the presence of a unit root at 5% significance level when the trend term was included on logged series only; (3) In this case, the ARIMA([1,4] , 1 , [1,4]) was slightly superior (Schwartz criterion) than the RW specification, but the F-test indicated that the hypothesis of all coefficients being zero could not be rejected at 1%.

Chapter 3

Temporal aggregation and revealed preference analysis of macro data: Are nonparametric tests biased towards nonrejection of low frequency data?

3.1 Introduction

Throughout this chapter we study from analytical and empirical standpoints how temporal aggregation affects Varian's nonparametric revealed preference analysis of consumption data. We are particularly interested in investigating whether and why the power of the GARP test seems to be positively correlated to the data frequency[46].

A couple of early studies[47] using the aforementioned framework on U.S. aggregate consumption data revealed that the test results were sensitive to data frequency. Recall from chapter 2 that Swofford and Whitney (1987) examined the

[46]Even though we are primarily concerned with previous empirical analyses of U.S. macroeconomic data, the first part of the discussion is purely theoretical and, therefore, applicable to the evaluation of data at all levels of aggregation. See Famulari (1995) for a recent application of Varian's framework to the analysis of household microeconomic (annual) data.

[47]Varian (1982), Manser and McDonald (1988) and Browning (1989) are additional examples of early works applying this approach to study aggregate consumption data. More recent contributions include Drake(1997) and Fleissig, Hall and Seater (2000).

separability structure of a representative consumer's utility function on goods and services, money holdings (assumed as a source of liquidity services) and leisure. They studied 15 years of U.S. quarterly data and found that consumption goods and leisure were weakly separable from liquid monetary assets in the representative utility function. In a subsequent paper, Swofford and Whitney (1988) expanded their analysis and compared results from quarterly and annual data; in contrast with tests results on quarterly data, annual data on goods and services, leisure and broad aggregates of monetary assets were found consistent with the existence of a well-behaved macro utility function rationalizing the data. The discrepancy may seem subtle: with quarterly data, the existence of such a macro utility function was mostly rejected, except when only relatively liquid monetary assets including small-time deposits were adopted as the relevant aggregate. Using data on broader aggregates of monetary assets led to the rejection of GARP-consistency in their first study. The authors argued that less liquid assets might be only partially adjusted within quarters, due to unobserved adjustment costs. However, a plausible alternative explanation – as we will show – is that the test may simply have failed to reject GARP-consistency on annual series due to its low power in low frequency datasets[48].

This second explanation finds some support also in empirical studies that do not include money holdings as a source of utility, revealing that the existence of short run monetary rigidities is a limited explanation for divergent results at different frequencies. FHS applied the same tests to subcategories of U.S. personal consumption expenditures, finding longer samples of GARP-consistent data with annual figures than with quarterly series[49]. Finally, the new evidence presented in chapter 1 is consistent with the alternative explanation, associating the power of the test with data frequency. Even

[48]The power of a test is its ability to reject a null hypothesis when it is indeed false; in other words, it is the probability of not committing a type II error (accept H0 when HA is true). Another issue, not examined here, is the size of the test, i.e., the probability of committing a type I error and reject H0 when it is true. Fleissig and Whitney (2003) studied, among other things, the size of the test if data contained measurement errors.

[49]This particular result was not sensitive to the alternative assumptions on consumers' expectations, as user costs of durable goods were calculated. See chapter 2 for details.

though long samples of both annual (1947-2000) and quarterly (1959:I-2000:IV) data were found to be consistent with GARP, our findings on the weak separability of subsets of data were sharply contrasting over annual and quarterly series. Many separability structures were rejected with quarterly data because the components of at least one of the supposedly separable subsets of goods contained GARP violations, whereas with annual data all structures passed both necessary and sufficient conditions (see section 1.4.2).

This chapter addresses, therefore, an aspect of the testing framework that has been overlooked in the literature. There is at least one fairly simple reason to believe that the use of low frequency data affects the power of the GARP test. Temporal aggregation can eliminate violations to the revealed preference axioms if conflicting choices in high-frequency data are averaged for the calculation of a single low-frequency observation (intra-period violations). Nevertheless, there are other less trivial aspects of temporal aggregation – involving choices made in distinct low-frequency intervals – that affect the test's ability to indicate the (in)existence of a well-behaved utility function rationalizing consumption data at some frequency. As we will demonstrate in the next section, a low-frequency perspective on the representative consumer's choices can remove many budget hyperplane intersections and, at the limit, even make finite samples of data uninformative for the GARP test.

The relevance of budget line intersections was best emphasized by Bronars (1987), who proposed ways to check the power of Varian's tests against alternative hypotheses of simulated irrational (random) behavior. Following recent contributions to this literature[50], Bronars' general approach is once more adopted in this chapter, and an additional attribute of typical consumption data is incorporated in the data simulation process: the nontrivial evolution of budget shares over time. To some extent, we will discuss how time-series aspects of original data may be an important step preceding data simulation and, consequently, the evaluation of the power of the test on a specific dataset.

[50]See Burton (1994), Cox(1997) and Mattei (2000) for applications and extensions to Bronars' approach.

The remainder of this chapter is organized in four sections. Section 3.2 presents a theoretical discussion of the possible impacts of time aggregation on the evaluation of GARP consistency, with graphical and numerical illustrations. Next we describe our simulation exercises, including Bronars's approach and extensions to it in section 3.3; the observation of some time series properties of the original data orient our discussion on potential shortcomings of Bronars' method in section 3.4. The final section contains our conclusions and some directions for future research.

Consistent with analytical findings, we will show that the estimated power of the GARP test is actually very high within quarterly datasets of U.S. consumption expenditures, but much lower with annual figures. The inclusion of leisure in the relevant set of commodities tends to raise the power of the test against some but not all of the alternative hypotheses; furthermore, we observe that some existing alternatives to Bronars' original simulation methods may generate quite misleading results under fairly common circumstances, supporting our view that future researchers may also benefit from an early investigation of the evolution of budget shares in actual data – preceding the interpretation of results from simulation exercises.

3.2 Temporal aggregation and GARP: the analytical perspective

In this section we formalize different aspects of temporal aggregation that explain how changes in the data frequency and the power of the GARP test are possibly associated. We start discussing and illustrating the relevance of budget intersections at different frequencies before specific circumstances are set out.

3.2.1 The relevance of budget intersections

Recall from our discussion in chapter 2 that a series of consumption choices passes the GARP test if the consumer's preference over sets of affordable bundles remains the same over time. Consider now what would happen if for all goods i=1,…,n, $[\mathbf{p^t x^t}/p^t_i] > [\mathbf{p^s x^s}/p^s_i]$, i.e., all affordable combinations of goods at time s were also affordable at t, but not a single combination of goods exhausting income at time t is available at time s. In other words, the budget hyperplanes representing all possible combinations of goods that exhaust incomes at time t and s do not intersect. Choices along the hyperplane available at time t are revealed preferred to any choice made at s, but since no bundle exhausting the budget at t is affordable at s, GARP cannot be violated, by definition. That raises a concern: the GARP test will not distinguish the behavior of a consumer that picks random points along such budget hyperplanes from that of a rational individual facing the same constraints. In Bronars'(1987) words, the dataset "contains no useful information about preference maximization" in that case. Before we formally demonstrate specific circumstances under which time aggregation eliminates budget intersections, a general illustration can help to clarify the issue.

Time aggregation, budget lines and GARP consistency: a first illustration. Suppose that a divorced couple of economists observe a series of choices made by their teenager daughter, who spends all her weekly allowance on two goods, say, video rentals and bags of candies. Her allowance varies according to her performance at school, but it is never less than $20 or more than $30/week. As prices change frequently, mom observes that her daughter faced the following series of budget constraints – choosing each time to purchase the combinations of goods in parentheses – over those weeks:

Week 1:	$4q_1 + 4q_2 = 24$	$(q_1^*=1; q_2^*=5)$
Week 2:	$4q_1 + 4q_2 = 24$	$(q_1^*=1; q_2^*=5)$
Week 3:	$6q_1 + 3q_2 = 24$	$(q_1^*=3; q_2^*=2)$
Week 4:	$2q_1 + 3q_2 = 30$	$(q_1^*=6; q_2^*=6)$

where q_1 and q_2 refer to numbers of video rentals and bags of candies, respectively. Over the last two weeks, both prices and her total expenditures changed, and she adjusted her choices in response. As mom plots the budget lines and choices actually made by her daughter over each week (figure $3.1^{(a)}$), she becomes extremely concerned and calls dad: "Our daughter must be changing… or maybe she is… not rational!" The source of concern is that over the third week her daughter purchased a combination of goods that had been affordable but not taken over the first two periods, rather than the typical bundle (repeatedly) purchased before, which was still feasible.

Hopeful that his daughter was, if anything, temporarily confused, dad decides to plot instead the biweekly average of quantities and prices over that month (figure $3.1^{(b)}$); in order to calm mom, he calls her back and states that there is no evidence to support the view that his daughter's choices were inconsistent with the utility maximization model at the biweekly frequency. However, after the call, he confesses to himself that the data at such frequency reveal no information at all about his daughter's rationality, because the corresponding budget lines do not intersect. There could be a well-behaved utility function rationalizing even random choices along those budget lines.

Although trivial, this illustration contains all elements that shall be formally (separately) discussed next: changes in relative prices, shifts in total expenditures and the averaging of high-frequency observations with temporal aggregation.

3.2.2 The GARP consistency of low frequency data: demonstrations

To show the different reasons why a dataset containing GARP violations at some high frequency can pass the GARP test within low frequency series, our strategy is setting out initially a general problem and then discussing classes of solutions, under particular circumstances. Following our formal discussion of particular classes of solutions, simple numerical exercises illustrate each of them. The reader must keep in mind throughout this discussion that the true frequency at which consumer choices are

made is unknown. The fundamental issue is a practical one, in the sense that we are investigating at which frequencies aggregate consumption flows are consistent with GARP, as well as whether our conclusions must necessarily be attributed to the separation of budget hyperplanes.

First we need to expand the notation used in chapter 2. Assume hereafter, for simplicity, that the highest frequency at which flows of consumption can be observed is semesterly[51], and that annual aggregates are calculated as the arithmetic averages of semesterly figures[52]. Let t(1) and t(2), s(1) and s(2) be the semesters of years T and S, respectively (see diagram 3.2). Superscripts define the time/period when a certain variable is observed, whereas subscripts will associate prices/quantities with specific goods. Let $\mathbf{p^T} \equiv [p_1{}^T \; p_2{}^T]$ be the (1 x 2) vector of annual prices for each of the n=2 goods in this simplified economy, where $p_i{}^T$ is a scalar calculated as the average of $p_i{}^{t(1)}$, $p_i{}^{t(2)}$, that is, the mean price of commodity i over year T. Similar notation is used for quantities, still generally represented by the letter x: $x_i{}^{t(.)}$ is a scalar representing the quantity of good "i" in a (2 x 1) vector (bundle) $\mathbf{x^{t(.)}}$; this semesterly vector of quantities $\mathbf{x^{t(.)}}$ can be thought of as some combination of goods that, if purchased, exhausts the disposable income over that period. The annual bundle $\mathbf{x^T}$, then, corresponds to a vector of average quantities allocated to the consumption of each good over year T. In equivalent notations:

$$\mathbf{x^T} \equiv \left\{ x^{t(1)} + x^{t(2)} \right\} \cdot 1/2 \equiv \left\{ \begin{bmatrix} x_1^{t(1)} \\ x_2^{t(1)} \end{bmatrix} + \begin{bmatrix} x_1^{t(2)} \\ x_2^{t(2)} \end{bmatrix} \right\} \cdot 1/2 \qquad \textbf{(3.1)}$$

The fundamental issue can be formalized as below:

---

[51] We adopted quarterly and annual frequencies in an earlier draft of this chapter (available upon request), in consistency with the data choices that will be made later for our empirical exercises; despite the slightly higher complexity, all analytical findings were unchanged and we opted for this simpler version.

[52] The Bureau of Economic Analysis (BEA) calculates annual figures on personal consumption expenditures – prices and real expenditures – by averaging quarterly figures, which are actually measured at annual rates.

Problem 1: *Let* $p_1^{t(1)}$, $p_2^{t(1)}$, $p_1^{s(1)}$, $p_2^{s(1)}$, $x_1^{t(1)}$, $x_2^{t(1)}$, $x_1^{s(1)}$, $x_2^{s(1)}$ *be the prices and quantities actually observed in the first semesters of distinct years* T *and* S, *such that a single GARP violation occurs within semesterly data:*

$$p_1^{t(1)} \cdot x_1^{t(1)} + p_2^{t(1)} \cdot x_2^{t(1)} \quad \geq \quad p_1^{t(1)} \cdot x_1^{s(1)} + p_2^{t(1)} \cdot x_2^{s(1)} \qquad \textbf{(3.2)}$$

$$p_1^{s(1)} \cdot x_1^{s(1)} + p_2^{s(1)} \cdot x_2^{s(1)} \quad > \quad p_1^{s(1)} \cdot x_1^{t(1)} + p_2^{s(1)} \cdot x_2^{t(1)} \qquad \textbf{(3.3)}$$

Are there positive numbers $p_i^{t(2)}$, $x_i^{t(2)}$, $p_i^{s(2)}$, $x_i^{s(2)}$ *for* i=1,2 *such that no GARP violation will be detected within annual data, as in expressions (3.4) and/or (3.5) below?*

$$\mathbf{p}^{\mathbf{T}} \cdot \mathbf{x}^{\mathbf{T}} \quad < \quad \mathbf{p}^{\mathbf{T}} \cdot \mathbf{x}^{\mathbf{S}} \qquad \textbf{(3.4)}$$

$$\mathbf{p}^{\mathbf{S}} \cdot \mathbf{x}^{\mathbf{S}} \quad < \quad \mathbf{p}^{\mathbf{S}} \cdot \mathbf{x}^{\mathbf{T}} \qquad \textbf{(3.5)}$$

Expression (3.2) implies that the bundle $\mathbf{x^{t(1)}} \equiv [x_1^{t(1)} \; x_2^{t(1)}]'$ is chosen at the first half of year T, while a second combination of goods $\mathbf{x^{s(1)}}$ is also affordable; (3.3) means that the choice is reversed at the first semester of year S, with $\mathbf{x^{s(1)}}$ being taken even though $\mathbf{x^{t(1)}}$ is still affordable; this pair of choices constitutes a GARP violation.

As we move to (3.4) and (3.5), though, it is important to notice first that we are setting out inequalities involving not only observed and, therefore, given bundles $\mathbf{x^{t(1)}}$ and $\mathbf{x^{s(1)}}$ and the corresponding vectors of prices at each of those semesters $\mathbf{p^{t(1)}} \equiv [p_1^{t(1)} \; p_2^{t(1)}]$ and $\mathbf{p^{s(1)}} \equiv [p_1^{s(1)} \; p_2^{s(1)}]$; those conditions involve *a priori* unknown vectors of prices and quantities for the remaining halves of each year. In fact, there will only be a solution to problem 1 if we can find positive values for those unknown variables so that either (3.4) or (3.5) holds – or both. Notice that in this first version of the problem we impose no restriction whatsoever on the total expenditures available at the second halves of those years; to find a solution, one can freely search for prices, quantities and total expenditures over those semesters so that (3.4) and/or (3.5) hold at the annual frequency.

The interpretation of inequalities (3.4) and (3.5) is quite the opposite of (3.2) and (3.3); either (3.4) or (3.5) holding means that no GARP violations can be detected with data on annual average prices and quantities for each good. Expression (3.4) alone implies that the annual bundle acquired in year S is not affordable when $\mathbf{x}^T$ is taken; consequently, one of the annual bundles is not revealed preferred to the other and such a pair of choices is necessarily GARP-consistent. Analogous reasoning applies to expression (3.5) holding alone, resulting in no annual GARP violation. Nevertheless, we still want to illustrate what it actually means having one or both expressions (3.4) and (3.5) true, a distinction that will be used in a coming subsection.

Points along and below the solid lines of figure 3.3 represent affordable bundles in each year – including the chosen bundles, marked with black dots; the dashed lines correspond to the cost of purchasing those same bundle but at different sets of prices[53]. In the first graph, identified as NV1 (no violation, case 1), all points along the solid lines for period T involve the same total expenditures, $\mathbf{p}^T\mathbf{x}^T$. The dashed line that also contains the bundle $\mathbf{x}^T$ is parallel to the continuous line for period S; therefore, it represents the cost of purchasing $\mathbf{x}^T$ when $\mathbf{p}^S$ is current; as this dashed line $\mathbf{p}^S\mathbf{x}^T$ is always higher than the continuous line for period S, $\mathbf{x}^T$ is not affordable when $\mathbf{x}^S$ is taken, precisely as stated in the inequality (3.5). Notice, on the other hand, that $\mathbf{p}^T\mathbf{x}^S$ is higher than the actual budget line for period T; thus, $\mathbf{x}^S$ is not affordable when $\mathbf{x}^T$ is chosen – confirming inequality (3.4). In other words, each of the observed choices is not directly revealed preferred to the other and, by definition, there is no GARP violation under those circumstances. In the second graph (NV2), however, only one of the bundles is not affordable at both times, which is still sufficient for GARP-consistency.

There is another way to see that both graphs in figure 3.3 represent choices consistent with the existence of well-behaved utility functions: one can draw convex indifference curves, tangent to the continuous budget lines, each containing one of the

[53]We use $\mathbf{x}^T$ for a particular bundle chosen as the set of prices $\mathbf{p}^T$ is observed, not any bundle affordable at T. Nevertheless, the budget line of some period T is simply referred to as $\mathbf{p}^T\mathbf{x}^T$, to avoid excessive notation.

black dots, so that consumers are shown to have picked points at the highest utility levels at each time.

Back to the discussion of this first and most general problem, it has a large number of solutions and we must focus on the existence of classes of solutions here. The first one is based on the fact previously discussed that GARP violations cannot occur if budget hyperplanes are separated. Another class of solutions does not require elimination of budget intersections and, as we will see, must have a very distinct interpretation. In all cases, however, our solutions will rely on the fact that arbitrary numbers for real expenditures, prices and, therefore, disposable income can be picked for the second semester of each year so that specific sets of inequalities hold.

a) Low frequency data and nonintersecting budget hyperplanes (Problem 2). Problem 1 is now restated so that our focus is on particular aspects of time aggregation with quite intuitive interpretations. All solutions to this next version of the problem will also solve Problem 1, but we will show later that they are not necessary conditions for a solution to the original one. It is convenient to reiterate that two annual budget hyperplanes will not intersect if $[\mathbf{p^T x^T}/p^T_i] > [\mathbf{p^S x^S}/p^S_i]$ for all goods i=1,2 . The new version is, then:

Problem 2: *Let* $p_1^{t(1)}$, $p_2^{t(1)}$, $p_1^{s(1)}$, $p_2^{s(1)}$, $x_1^{t(1)}$, $x_2^{t(1)}$, $x_1^{s(1)}$, $x_2^{s(1)}$ *be the prices and quantities actually observed in the first semesters of distinct years* T *and* S, *such that a single GARP violation occurs within semesterly data (precisely as in Problem 1):*

$$p_1^{t(1)} \cdot x_1^{t(1)} + p_2^{t(1)} \cdot x_2^{t(1)} \quad \geq \quad p_1^{t(1)} \cdot x_1^{s(1)} + p_2^{t(1)} \cdot x_2^{s(1)} \qquad \textbf{(3.2)}$$

$$p_1^{s(1)} \cdot x_1^{s(1)} + p_2^{s(1)} \cdot x_2^{s(1)} \quad > \quad p_1^{s(1)} \cdot x_1^{t(1)} + p_2^{s(1)} \cdot x_2^{t(1)} \qquad \textbf{(3.3)}$$

Are there positive numbers $p_i^{t(2)}$, $x_i^{t(2)}$, $p_i^{s(2)}$, $x_i^{s(2)}$ *for* i=1,2 *such that the* <u>*annual*</u> *budget lines for years* T *and* S *do not intersect, as in either one of the series of inequalities below?*

$$\mathbf{p^T x^T} \;/\; p_i^T \;>\; \mathbf{p^S x^S} / p_i^S \qquad \text{for } i = 1,2 \qquad \textbf{(3.6)}$$

$$\mathbf{p^S x^S} \;/\; p_i^S \;>\; \mathbf{p^T x^T} / p_i^T \qquad \text{for } i = 1,2 \qquad \textbf{(3.7)}$$

Finding values for 8 unknowns such that a pair of inequalities is satisfied, either (3.6) or (3.7) depending on which budget line is higher, may seem to be a bit harder than the task involved in the search for a solution to Problem 1. However, by solving this version of the problem, we can stress that low-frequency budget lines may end up not intersecting due to clearly distinct aspects of time aggregation, as we describe next[54].

1st set of solutions to problem 2: relative price smoothing (RPS). It is not hard to find positive prices of the two goods over the second semesters of years T and S so that the two annual budget lines become parallel, i.e., that the annual relative prices are unchanged for those years and $p_1^T/p_2^T = p_1^S/p_2^S$; in more detailed notation, such solution requires that:

$$p_1^T/p_2^T \equiv \frac{p_1^{t(1)} + p_1^{t(2)}}{p_2^{t(1)} + p_2^{t(2)}} = \frac{p_1^{s(1)} + p_1^{s(2)}}{p_2^{s(1)} + p_2^{s(2)}} \equiv p_1^S/p_2^S \qquad \textbf{(3.8)}$$

Recall that $p_1^{t(1)}$, $p_2^{t(1)}$, $p_1^{s(1)}$, $p_2^{s(1)}$ are parameters in all versions of the problem; the following are sufficient (not necessary) conditions for a solution to problem 2: $p_1^{t(2)} = p_1^{s(1)}$, $p_2^{t(2)} = p_2^{s(1)}$, $p_1^{s(2)} = p_1^{t(1)}$, $p_2^{s(2)} = p_2^{t(1)}$. Such choice of values implies that the budget

[54]Over the next few pages, we propose particular solutions to the mathematical problems basically showing prices and/or quantities shifts that lead to specific results; we do not provide economic explanations for the ultimate causes of those changes in each case, but one can think of exogenous supply shocks to motivate all price changes and income shocks as the ultimate cause for shifts in total expenditures.

line for the second semester of year T is parallel to the one for the first semester of S, whereas the line for the second semester of year S is parallel to the one for the first half of year T, as illustrated in figure 3.4[(a)]. The black dots in that figure represent the semesterly GARP violations, involving the choices over the first half of each year. As we substitute the values above on both sides of (3.8), we get:

$$p_1^T/p_2^T = \frac{p_1^{t(1)} + p_1^{s(1)}}{p_2^{t(1)} + p_2^{s(1)}} = \frac{p_1^{t(1)} + p_1^{s(1)}}{p_2^{t(1)} + p_2^{s(1)}} = p_1^S/p_2^S \qquad \textbf{(3.9)}$$

For any combination of goods exhausting income over the second halves of each year, the annual bundles will be points along parallel annual budget lines, as represented in figure 3.4[(b)].

Another combination of arbitrarily chosen prices that makes expression (3.8) true is: $p_1^{t(2)} = p_2^{t(1)}$; $p_2^{t(2)} = p_1^{t(1)}$; $p_1^{s(2)} = p_2^{s(1)}$; $p_2^{s(2)} = p_1^{s(1)}$; in this case, (3.8) becomes:

$$\frac{p_1^{t(1)} + p_2^{t(1)}}{p_2^{t(1)} + p_1^{t(1)}} = \frac{p_1^{s(1)} + p_2^{s(1)}}{p_2^{s(1)} + p_1^{s(1)}} = 1 \qquad \textbf{(3.10)}$$

Regardless of which set of arbitrary prices above is picked, the resulting vectors of annual prices become parallel, meaning that from an annual perspective the relative prices are in fact unchanged. Those solutions do not demand any additional condition regarding the quantities of each good purchased throughout those years – besides obviously the exhaustion of all available income at each quarter; suppose for example that random combinations of goods are picked along the budget lines for the last semesters of T and S; if at the annual basis the maximum affordable amount of some good "i" in year T ($\mathbf{p^T x^T}/p_i^T$) turns out to be smaller than in S ($\mathbf{p^S x^S}/p_i^S$), the inequalities in (3.7) hold because budget lines are parallel and the consumer can also afford a smaller

amount of the second good in year T than in S. (3.6) holds in the opposite case, using the same reasoning.[55]

In sum, if high-frequency changes in relative prices are offset in the calculation of the annual average, the low-frequency series become uninformative for GARP-consistency evaluations. We refer to this aspect of time aggregation as relative price smoothing (RPS) appealing to the fact that relative prices in finite samples of high-frequency data may behave as white noise processes, subject to small temporary shocks; as one aggregates consumption flows into lower frequency datasets, those shocks are averaged and may become insignificant.

2nd set of solutions to problem 2: real expenditure shifting (RES). Another class of solutions for this version of the problem involves finding values for real expenditures (quantities) over the second halves of each year so that one of the annual budget lines can be shown to have higher intercepts than the other. In order to obtain a sufficient solution, eventual changes in relative prices – along with the changes in expenditures – over those periods must also be considered[56].

In words, the idea is to set up annual budget constraints so that the upper bound for quantities of the least expensive good in year S is, by definition, smaller than the upper bound for quantities of the most expensive good acquired over year T. After that, it is only a matter of rearranging expressions to show $x_i^{t(2)}$ and $x_i^{s(2)}$ as functions of all other variables. Figure 3.5 illustrates a couple of cases in which this method is applied, the distinction being particular evolutions of relative prices; we present more than one example simply because the upper bounds aforementioned depend not only on the

[55]The pair of annual budget lines could also coincide, if annual average expenditures were the same over the two years. In such case, budget hyperplane separation would not occur, but the main conclusion would still hold: two bundles along a same budget line will never constitute a GARP violation, as one can deduce from the strict inequality signal in the definition of GARP.

[56]Similarly to the same way we identified price vectors that were sufficient for the first set of solutions (regardless of specific combinations of goods actually chosen), we want to show now that for any evolution of prices over those periods, one can obtain – from a general numerical expression – arbitrary values for quantities on the second semesters of those years that guarantee nonintersecting annual budget lines.

disposable income available at each time, but on possibly changing relative prices, as we discuss next.

Assume once more that the bundles chosen in the first semesters of each year satisfy expressions (3.2) and (3.3) – a semesterly GARP violation, as represented by the black dots on the left-hand-side graphs of figure 3.5. For any price vectors that turn out to be current at the second semesters of each year, $\mathbf{p^T}$ and $\mathbf{p^S}$ are still calculated as the averages of semesterly prices; let p^T_{max}, p^S_{max} and p^T_{min}, p^S_{min} be scalars representing the highest and lowest annual prices in each year, respectively; (3.11) and (3.12) below show sufficient conditions involving values for $x_i^{t(2)}$ and $x_i^{s(2)}$ (i=1,2) and some arbitrary numbers α,β (α≥β>0) so that the inequalities in (3.6) will hold:

$$x_i^T \equiv (x_i^{t(1)} + x_i^{t(2)}) \cdot 1/2 = \alpha \cdot \frac{p^T_{max}}{np_i^T} \qquad \text{for i=1,2} \quad \text{(n=2 goods)} \qquad \mathbf{(3.11)}$$

$$x_i^S \equiv (x_i^{s(1)} + x_i^{s(2)}) \cdot 1/2 = \beta \cdot \frac{p^S_{min}}{np_i^S} \qquad \text{for i=1,2} \quad \text{(n=2 goods)} \qquad \mathbf{(3.12)}$$

The right-hand-sides in the expressions above are values particularly chosen in order to make annual total expenditures multiples of specific prices at each year: $\mathbf{p^T x^T} \equiv p_1^T x_1^T + p_2^T x_2^T = \alpha\, p^T_{max}$ and $\mathbf{p^S x^S} \equiv p_1^S x_1^S + p_2^S x_2^S = \beta\, p^S_{min}$; consequently, α constitutes the largest amount of the most expensive good purchased in year T, whereas β is the largest amount of any good purchased in year S. Substituting $\mathbf{p^T x^T} = \alpha p^T_{max}$ and $\mathbf{p^S x^S} = \beta p^S_{min}$ in (3.6), the inequalities will hold because the right hand sides will be either larger than or equal to α, whereas the left hand sides will be, at most, equal to β (and smaller than α by definition).

To visualize the solution, take initially figures $3.5^{(a)}$ and $3.5^{(b)}$; notice we omitted the budget lines for the second semester of each year to keep the interpretation of figures as straightforward as possible. In this first case budget lines at both frequencies reveal that good 1 is the most expensive good purchased in year S – vertical intercepts are further distant from the origin than horizontal ones –, with good 2 becoming relatively more expensive in year T. Once more, the largest amount of the most expensive good

possibly purchased over year T is α units of good 2, which by definition is bigger than the largest amount of any good purchased in year S, at most β units of that same good. In the second pair of graphs (3.5$^{(c)}$ and 3.5$^{(d)}$), good 1 is at each and every period the most expensive commodity; still, the largest possible amount of good 1 purchased in year T is α units, which is bigger than the largest amount of any good possibly purchased in S (β units of good 2).

Finally, one can set β as a fraction of α [say, β=(0.90α)] and rearrange (3.11) and (3.12), so that they are turned into conditions that allocations over the second semesters of each year must pass in order to guarantee the separation of annual budget lines:

$$x_i^{t(2)} = 2 \cdot \alpha \cdot \frac{p_{max}^{T}}{np_i^{T}} - x_i^{t(1)} \qquad \textbf{(3.13)}$$

$$x_i^{s(2)} = 2 \cdot (0.90 \times \alpha) \cdot \frac{p_{min}^{S}}{np_i^{S}} - x_i^{s(1)} \qquad \textbf{(3.14)}$$

In practice, a particular solution requires additionally that α is made large enough so that the right-hand sides of (3.13) and (3.14) are strictly positive, avoiding negative quantities in the second halves of those year. One can see that the problem has multiple solutions because the parameters α,β can assume infinite values: the smaller fraction β is relatively to α, the budget lines for years T and S move further apart.

In sum, the interpretation for this second set of solutions is that, regardless of changes in relative prices, annual budget lines may not intersect if real expenditures have very different magnitudes over each of those years. Our emphasis was on the fact that no restriction on the evolution of prices must be imposed *a priori* for a solution; the numerical expressions above generate solutions – through what we named the real expenditure shifting (RES) effect – that incorporate any changes in relative prices.

b) GARP consistency of annual data with intersecting budget hyperplanes. Datasets containing semesterly violations may be GARP-consistent at the annual frequency even if annual budget hyperplanes intersect and, therefore, low frequency data are not

uninformative for the GARP test. To show this, recall the distinction between two cases of GARP-consistent choices from intersecting budget lines, NV1 and NV2 (figure 3.3); we focus on the most stringent[57] one (NV1) to reformulate our original problem:

***Problem 3**: Let* $p_1^{t(1)}$, $p_2^{t(1)}$, $p_1^{s(1)}$, $p_2^{s(1)}$, $x_1^{t(1)}$, $x_2^{t(1)}$, $x_1^{s(1)}$, $x_2^{s(1)}$ *be the prices and quantities actually observed in the first semesters of distinct years* T *and* S, *such that a single GARP violation occurs within semesterly data (precisely as in Problems 1 and 2):*

$$p_1^{t(1)} \cdot x_1^{t(1)} + p_2^{t(1)} \cdot x_2^{t(1)} \quad \geq \quad p_1^{t(1)} \cdot x_1^{s(1)} + p_2^{t(1)} \cdot x_2^{s(1)} \qquad \textbf{(3.2)}$$

$$p_1^{s(1)} \cdot x_1^{s(1)} + p_2^{s(1)} \cdot x_2^{s(1)} \quad > \quad p_1^{s(1)} \cdot x_1^{t(1)} + p_2^{s(1)} \cdot x_2^{t(1)} \qquad \textbf{(3.3)}$$

Are there positive numbers $p_i^{t(2)}$, $x_i^{t(2)}$, $p_i^{s(2)}$, $x_i^{s(2)}$ *for* i=1,2 *such that annual budget lines do intersect – either expression (3.15) or (3.16) below holding – and also that each of the annual bundles are not directly revealed preferred to the other, as in (3.4) and (3.5)?*

$$\mathbf{p^T x^T} \;/\; p^T{}_{max} \;<\; \mathbf{p^S x^S} \;/\; p^S{}_{min} \qquad \textbf{(3.15)}$$

$$\mathbf{p^S x^S} \;/\; p^S{}_{max} \;<\; \mathbf{p^T x^T} \;/\; p^T{}_{min} \qquad \textbf{(3.16)}$$

$$\mathbf{p^T x^T} \;<\; \mathbf{p^T x^S} \qquad \textbf{(3.4)}$$

$$\mathbf{p^S x^S} \;<\; \mathbf{p^S x^T} \qquad \textbf{(3.5)}$$

This time we will not rely on the two aspects leading to the elimination of budget intersections: the RPS and RES effects. Suppose first that the representative consumer faces the same set of prices throughout each year, so that annual budget lines will have the same slope of the two semesterly ones; time aggregation, in this case, does not smooth intra-period changes in relative prices simply because they are fixed. Also assume that the consumer allocates the same total expenditure on each semester of a

[57]NV2 can be seen as a special case of NV1, as NV2 involves only a subset of the conditions for NV1.

given year, which implies that semesterly budget lines are not only parallel to each other and to the annual line: they actually coincide. Since the same total expenditure is spent at each semester, time aggregation will not make GARP violations less likely via the "shifting" effect on the transition from semesterly to annual budget lines. Semesterly bundles purchased over any given year can be represented as points along the same line, as well as their average – the annual bundle. The aforementioned conditions are summarized below:

$$\mathbf{p}^{t(1)} = \mathbf{p}^{t(2)} = \mathbf{p}^{T} \neq \mathbf{p}^{S} = \mathbf{p}^{s(1)} = \mathbf{p}^{s(2)} \qquad \textbf{(3.17)}$$

$$\mathbf{p}^{s(1)}\mathbf{x}^{s(1)} = \mathbf{p}^{s(2)}\mathbf{x}^{s(2)} = \mathbf{p}^{S}\mathbf{x}^{S} \;\; ; \;\; \mathbf{p}^{t(1)}\mathbf{x}^{t(1)} = \mathbf{p}^{t(2)}\mathbf{x}^{t(2)} = \mathbf{p}^{T}\mathbf{x}^{T} \qquad \textbf{(3.18)}$$

As semesterly and annual budget lines coincide, it is true that the representative consumer is expected to make the same choices at each semester of a given year; however, we explicitly assume, for the sake of this demonstration, that his expenditures at the first semester of a pair of years are indeed abnormal, revealing a GARP violation[58].

Along with (3.2) and (3.3), the expressions above imply that the annual budget lines generally referred to as $\mathbf{p}^{T}\mathbf{x}^{T}$ and $\mathbf{p}^{S}\mathbf{x}^{S}$ will intersect, since (3.17) and (3.18) actually mean that they coincide with budget lines for the first semesters containing a GARP violation. Therefore, either expression (3.15) or (3.16) is always satisfied, regardless of specific combinations of quantities possibly observed along the budget lines for the second semesters of each year. Nevertheless, it remains the task of finding

[58]Assuming a "temporary" irrationality may seem unappealing at first, and we thank Dr. Walter Thurman for pointing that out. However, there are reasons why such abnormality may occur; one could say that even though choices are observable at the semesterly frequency, the consumers actually maximize the aggregate flow of goods for a whole year – with high frequency changes in that flow being meaningless. Also, the existence of measurement errors associated with the timing of sales reporting in specific sectors can also create the impression that the composition of bundles change along the year. Rather than a theoretical possibility, this last issue has been explicitly addressed by Wilcox (1992), as discussed in chapter 1.

quantities along those budget lines so that expressions (3.4), (3.5) and (3.18) are satisfied.

Figure 3.6 illustrates the occurrence of a semesterly violation under the specific circumstances described above – the left-hand side graph – and also what the solution at the annual basis must look like.

Let $\mathbf{x^A}$ be the point of intersection of budget lines, i.e., the unique vector of quantities such that $\mathbf{p^T x^T = p^T x^A}$ and $\mathbf{p^S x^S = p^S x^A}$. In both graphs of figure 3.6, such point is represented by a white dot, contrasting to the black dots characterizing the bundles actually purchased at each period. Focusing on the second graph (Annual NV1), notice that the intersection point divides the budget line for year T in two segments; let T^u be the locus of points along the budget line $\mathbf{p^T x^T}$ that are not affordable in year S, whereas T^d is the segment of $\mathbf{p^T x^T}$ containing bundles that are inside the set of choices affordable in year S – $\mathbf{x^{t(1)}}$ and $\mathbf{x^A}$ itself included in this second group. In the same fashion the budget line $\mathbf{p^S x^S}$ can be divided in S^u and S^d, representing sets of bundles not affordable / also affordable in year T, respectively. In our 2-good economy, it is easy and convenient to set out equivalent but more formal definitions for those sets. Assume without loss of generality that the budget lines for year T are steeper that the ones for year S, i.e., $p_1^{t(1)}/ p_2^{t(1)} > p_1^{s(1)}/ p_2^{s(1)}$, precisely as illustrated in NV1. The new definitions are:

Definitions:

T^u: $\{\mathbf{x^{t(.)}} \in \mathbf{p^T x^T} \mid x_1^{t(.)} < x_1^A \text{ and } x_2^{t(.)} > x_2^A\}$;

T^d: $\{\mathbf{x^{t(.)}} \in \mathbf{p^T x^T} \mid x_1^{t(.)} \geq x_1^A \text{ and } x_2^{t(.)} \leq x_2^A\}$;

S^u: $\{\mathbf{x^{s(.)}} \in \mathbf{p^S x^S} \mid x_1^{s(.)} < x_1^A \text{ and } x_2^{s(.)} > x_2^A\}$;

S^d: $\{\mathbf{x^{s(.)}} \in \mathbf{p^S x^S} \mid x_1^{s(.)} \geq x_1^A \text{ and } x_2^{s(.)} \leq x_2^A\}$.

We can now state the conditions for the existence of solutions to problem 3 under the stricter circumstances described above (no relative price smoothing or shifting effect):

Proposition 1: There exist solutions to problem 3 that also satisfy expressions (3.17) and (3.18) if and only if there are vectors of positive quantities $\mathbf{x^{t(2)}}$, $\mathbf{x^{s(2)}}$ such that $\mathbf{x^T} \in T^u$ and $\mathbf{x^S} \in S^u$.

Proof: Suppose otherwise; take any set of positive scalars $x_i^{t(2)}$, $x_i^{s(2)}$ for i=1,2 such that $\mathbf{x^T} \in \mathbf{p^T x^T}$, $\mathbf{x^S} \in \mathbf{p^S x^S}$ and, therefore, that expression (3.18) holds; as T^d and T^u are partitions of the line $\mathbf{p^T x^T}$, it cannot be true that $\mathbf{x^T} \in T^u$ and $\mathbf{x^T} \in T^d$; $\mathbf{x^T}$ cannot be an element in both sets. If $\mathbf{x^T} \notin T^u$, then $\mathbf{x^T} \in T^d$ and $x_1^T \equiv [(x_1^{t(1)}+x_1^{t(2)})/2] \geq x_1^A$ and $x_2^T \equiv [(x_2^{t(1)}+x_2^{t(2)})/2] \leq x_2^A$; then, it is also true that $p^S x^T < p^S x^A$; since $p^S x^A = p^S x^S$ by construction, $p^S x^T < p^S x^S$ and expression (3.5) is violated, as in NV1. The same reasoning applies to bundles chosen in the second semester of S, creating a violation of (3.4). Thus, $\mathbf{x^T} \in T^u$ and $\mathbf{x^S} \in S^u$ are necessary conditions for a solution of problem 3 also satisfying (3.17) and (3.18). Showing that those conditions are also sufficient is straightforward: $\mathbf{x^T} \in T^u$ implies that $\mathbf{p^S x^T > p^S x^A}$ and, since $\mathbf{p^S x^A = p^S x^S}$, inequality (3.5) is satisfied; again, same reason applies for $\mathbf{x^S} \in S^u$.

Intuitively, if the bundles chosen in the second half of year T do not "pull" the annual average out the set of choices also affordable throughout year S, $\mathbf{x^T}$ is still revealed preferred to choices along $\mathbf{p^S x^S}$, violating one of the conditions for NV1; similarly, if the annual (average) bundle for year S does not lie outside the set of choices also affordable throughout year T, $\mathbf{x^S}$ is revealed preferred to choices along $\mathbf{p^T x^T}$, also violating a condition for NV1. As for sufficiency, the average bundles being points along one of the lines but beyond the other year's budget set is precisely equivalent to inequalities (3.4), (3.5) and (3.18) holding[59].

Figure 3.7 illustrates a second application of this sort of solution, without RPS or RES effects. The difference from the previous illustration is that the single common

[59]Since budget lines are linear combinations of quantities and the average of coordinates for two points along some line defines another point along that same line, in practice proposition 1 implies that there will always be a solution to this last problem as long as budget intersections are distant enough from intercepts such that $\mathbf{p^{t(.)}x^{t(.)}}/p_2^{t(.)} > 2x_2^A$, $\mathbf{p^{t(.)}x^{t(.)}}/p_1^{t(.)} < 2x_1^A$, $\mathbf{p^{s(.)}x^{s(.)}}/p_1^{s(.)} > 2x_1^A$ and $\mathbf{p^{s(.)}x^{s(.)}}/p_2^{s(.)} < 2x_2^A$.

bundle exhausting income in the first semesters of those year is actually taken. Notice that this time we also plotted the choices for the second semesters of those years (on the left-hand-side graph); recalling that the average of coordinates for two points along a given line is, itself, a third point on that line, the annual bundles were "forced" to lie outside of the T^d and S^d segments and annual choices are once more GARP-consistent.

We finally illustrate (without proving) that problem 3 may not have a solution under special circumstances. All conditions in proposition 1 are implicitly adopted in this last graph (figure 3.8), but the budget intersection occurs much closer to one of the axes now. The black dots once more represent observed GARP-violating choices over first semesters of the two years; notice that all combinations of goods exhausting disposable income in one of the years are directly revealed preferred to the bundle actually taken in one of the semesters of the other period. This semesterly violation might still be "reversed" at the annual frequency by subsequent choices in each of those years, but not likely into a NV1 case: the GARP-violating choices are rather extreme, in the sense that the consumer spends most of his income on x_1 at the first semester of year T and almost none of it in the beginning of year S. Even if he spends its entire disposable income on x_2 in the second half of year T, the annual average bundle $\mathbf{x}^{\mathbf{T}}$ along that same budget line will still lie below the intersection point $\mathbf{x}^{\mathbf{A}}$ (white dot), which means the average of choices made in year S ($\mathbf{x}^{\mathbf{S}}$) would still be revealed preferred to $\mathbf{x}^{\mathbf{T}}$. Notice that the choice made in the second half of S could still imply no GARP violation at the annual frequency, as the average bundle $\mathbf{x}^{\mathbf{S}}$ might be "pulled off" to the right of the budget intersection; that means NV2 clearly seems possible, but not NV1.

3.2.3 Numerical examples illustrating the three main analytical findings

We now illustrate with numbers both classes of solutions to problem 2, as well as an example of problem 3 that indeed contains a solution. In all cases, we will assume given prices and quantities of two commodities in the first semesters of years T and S;

those constitute a same GARP violation at the semesterly frequency, as set out in the initial statements of the problems. Then we will essentially pick arbitrary prices and/or quantities for the remaining semesters – using the algebraic expressions derived before – that will lead to GARP consistent choices at the annual frequency.

The first example involves choosing only prices for the second semesters of those years so that, as we calculate average prices throughout the periods, the two annual budget lines are parallel – 1st set of solutions, problem 2. To avoid coinciding annual budget lines, we assume that total expenditures are the same over each semester of a given year but not across different years. Table 3.1 presents the (given) prices and quantities for the first semesters of each year, the calculated prices for the last semesters and the annual averages.

Notice first that choices made over the first semesters indeed violate GARP: the consumer could have purchased 200 units of good 1 and 250 units of good 2 at s(1) – as he did in t(1) – spending $32,500.00, less than the $35,000.00 that he actually spent at that time; in other words, the choice made at t(1) was also affordable but not taken at s(1). On the other hand, he could have purchased 100 units of good 1 and 500 units of good 2 at t(1) spending the same $30,000.00 that he actually used in that semester. Therefore, a different ordering of preferences must have motivated his choices at each time.

To calculate prices over the remaining semesters in each year, we use the price settings previously discussed: $p_1^{t(2)} = p_1^{s(1)}$, $p_2^{t(2)} = p_2^{s(1)}$, $p_1^{s(2)} = p_1^{t(1)}$, $p_2^{s(2)} = p_2^{t(1)}$; such choice implies that:

$$p_1^T/p_2^T = \frac{p_1^{t(1)} + p_1^{s(1)}}{p_2^{t(1)} + p_2^{s(1)}} = \frac{p_1^{t(1)} + p_1^{s(1)}}{p_2^{t(1)} + p_2^{s(1)}} = p_1^S/p_2^S \qquad \textbf{(3.9)}$$

The price of good 1 ended up being the same in years T and S, but the price of good 2 was reduced at s(2), rather than increased as in the second half of year T, t(2). If T and S were subsequent years, the higher price of good 2 lasting for a couple of

semesters could be seen as a temporary shock, affecting annual statistics symmetrically: Annual prices became precisely the same over those years, implying that the annual budget lines were parallel. Furthermore, since total expenditures were assumed always higher in year S, any set of feasible bundles exhausting all disposable income at the second half of T and S would imply a GARP consistent pair of annual choices – making the task of finding specific quantities in this solution absolutely dispensable.

Recall that the second class of solutions to problem 2 involves mainly setting convenient numbers for quantities over second semesters of years T an S as functions of all prices, which will consequently be treated as parameters of the problem. For simplicity, we assume that prices fluctuate around the values observed in the beginning of each year, due to arbitrarily small exogenous shocks. To calculate quantities, we rely on a pair of expressions derived before:

$$x_i^{t(2)} = 2 \cdot \alpha \cdot \frac{p_{max}^{T}}{np_i^{T}} - x_i^{t(1)} \qquad \textbf{(3.13)}$$

$$x_i^{s(2)} = 2 \cdot (0.90 \times \alpha) \cdot \frac{p_{min}^{S}}{np_i^{S}} - x_i^{s(1)} \qquad \textbf{(3.14)}$$

Recall that α is arbitrarily picked so that all quantities are positive; we set α=300. Table 3.2 shows once more the GARP-violating prices and quantities for the first semesters of each year, given prices for both goods in all semesters and, finally, the calculated quantities that guarantee budget hyperplane separation – these last numbers being presented in bold italic, for easier identification.

As discussed before, our choice of quantities was not limited by any sort of predetermined total expenditure at the second halves of each year. The relatively larger amount of good 1 purchased at the first semester of year S – at a price higher than the current one at t(1) – was compensated by significantly smaller figures in the subsequent semester, bringing the annual average down. To make sure that the annual averages presented in table 3.2 indeed imply budget hyperplane separation, we also calculated the maximum amounts of each good that could be affordable at the annual basis, i.e., the

budget line intercepts. If the consumer decides to use all his annual resources to purchase only one of the goods, he is able to afford 300 units of goods 1 and 675 units of good 2, in year T. Making the same "extreme" choices with the resources available in year S, the maximum amounts of those goods would be 270 and 607, respectively.

Starting one last time with GARP-violating choices for the first semesters of years T and S, we finally illustrate a case of problem 3 that indeed has a solution. We assume now that prices and total expenditures are fixed along the semesters of a given year (table 3.3). One simply needs to find feasible bundles for the remaining quarters of those years that will make each of the annual choices not directly revealed preferred to the other; in other words, the annual bundle purchased in T must be unaffordable in S, and vice-versa.

The solution is trivially obtained by setting extreme allocations of resources for the last semesters of each year: the consumer spends all his disposable income on good 1 at s(2) and the very opposite allocation of resources is made over the second half of year T. As annual bundles are calculated, $\mathbf{x}^{\mathbf{T}}$ (100 units of good 1, 500 units of good 2) is not affordable at the average prices of year S; the consumer would have to spend $35,000.00, which exceeds the $33,750.00 available at that time; therefore, x^T is not revealed preferred to x^S. On the other hand, the combination of goods purchased in S (225 units of good 1, 225 units of good 2) is also not affordable at year T, as it would cost the consumer $31,500.00, beyond his $30,000.00 budget in year T.

The simple averaging of data due to the aggregation of consumption flows into annual figures was sufficient to eliminate the semesterly GARP violation, even though annual budget lines still intersect. The intuition is trivial: the annual bundles for each year ended up being outside the budget set for the other year.

3.2.4 Solving one last puzzle: can time aggregation create GARP violations?

So far we have investigated how GARP violations can be observed at a relatively higher frequency but not when the level of temporal aggregation increases. Now we illustrate the opposite possibility: low frequency data may contain GARP violations even though they are not detected at higher frequencies. To show that, we will refer back to the young lady making choices on how to spend her weekly allowance, the example that motivated this whole section.

The story is mostly the same: the two economists' daughter spends her exogenously set income on movie rentals and candies; her choices are unchanged over the first two weeks, but when total income and prices change in the second half of the month she makes very distinct allocations of resources. Mom as before observes choices weekly – described below – and dad relies on a biweekly accounting (figure 3.9):

Week 1:	$4q_1 + 4q_2 = 24$	$(q_1^*=1; q_2^*=5)$
Week 2:	$4q_1 + 4q_2 = 24$	$(q_1^*=1; q_2^*=5)$
Week 3:	$15q_1 + 5q_2 = 30$	$(q_1^*=2; q_2^*=0)$
Week 4:	$5q_1 + 2.5q_2 = 25$	$(q_1^*=3.5; q_2^*=3)$

The left-hand-side graph in figure 3.9 shows that mom is not able to detect any violation to the general axiom of revealed preference analysis, even though weekly budget lines do intersect often. However, when dad calculates average prices and quantities on a biweekly basis (figure 3.9(b)), his daughter's choices are indeed inconsistent with GARP: the average bundles are directly revealed preferred to each other, meaning that even though they were affordable in both occasions, the consumer's choices expressed different orderings of preferences each time.

Due to the particularity of the conditions involved in this last case, we were not able to establish a general expression that could guarantee the result, given some initial

high-frequency choices containing no GARP violation. Nevertheless, we will keep in mind the existence of this possibility, especially when our simulation exercises are run.

3.2.5 Summary of analytical findings and their interpretations

In this section we have shown different reasons why GARP violations observed at a relatively higher frequency (semesterly) may go undetected as consumption flows are aggregated on an annual basis with standard procedures (simple average). We also saw that the opposite case is possible: under particular circumstances, there may be GARP violations in low frequency data calculated from GARP-consistent choices at higher frequencies. It is useful to summarize those analytical findings and their interpretations before we move on to the empirical analysis.

First, two aspects possibly affecting test results through the budget separation mechanism were formally established, as classes of solutions to problem 2; we refer to them as the real expenditure shifting effect (RES) and the relative price smoothing effect (RPS). The occurrence of large income shifts – consistent with the commonly observed growth of real consumption expenditures over recent decades – were shown to be a plausible explanation for budget intersections at some but not all frequencies. One can speculate that the RES effect becomes particularly important if data on quantities or real expenditures change monotonically, so that budget hyperplanes tend to shift steadily over the studied sample. On the other hand, our discussion on the RPS effect seems to indicate that if high frequency shocks to relative prices are small and transitory – possibly being reversed during a same calendar year – low-frequency budget hyperplanes for two years may not intersect, even if a couple of their quarterly ones do and income shifts are relatively small. In any case, rather than pursuing exploratory time series analyses on the particular relevance of each of these two initial factors, we will study their joint impacts on the power of the GARP test using Bronars' (1987)

simulation approach; both original contribution and its extensions are the scope of the next section.

The third analytical finding – presented as the solution to problem 3 – was due to the fact that, as a couple of high-frequency inconsistent choices are averaged with other consistent ones, pairwise comparisons of averaged observations may indeed reveal no GARP violation whatsoever. Obtained in the absence of the RES and RPS effects, such a result has interesting interpretations. First, notice that rather than failing to detect possibly inconsistent choices with nonintersecting budget lines, the test in this case is properly pointing out the GARP consistency of averaged figures, in intersecting (annual) budget lines. Consequently, the elimination of GARP violations in a low frequency dataset may also be (at least partially) attributed to reasons that have nothing to do with the power of the test; in other words, GARP violations can be undetectable with low frequency data for multiple reasons, not exclusively due to nonintersecting budget hyperplanes. On theoretical grounds this last case seems to provide some support for Swofford and Whitney's (1988) allegation, according to which GARP violations occurring only within high-frequency data actually reflect unobservable adjustment costs in the short run[60]; for example, if relative prices change and consumers do face significant costs to adjust their bundles in the short run, a couple of quarters may pass before the new optimal allocation of resources is achieved. But since the GARP test is based on a frictionless world, results would likely indicate GARP violations at quarterly/semesterly but possibly not at the annual frequency, as in the solution for problem 3.

Therefore, the solutions for problems 2 and 3 can be considered the formal expressions of nonexclusive explanations discussed in our introduction: GARP violations may be detected within quarterly and not annual datasets either because budget intersections are less likely at lower frequencies or because high-frequency

[60]The authors were mainly concerned about adjustment of financial asset holdings, which were assumed to provide utility flows from liquidity services. Even though this paper does not discuss money or any other financial asset in utility functions, the same logic applies to adjustments of the stock of consumer durables.

adjustments are costly, or even both – as long as they apply to different pairs of observations in the same sample[61]. Due to the nature of the Swofford and Whitney's argument, involving unobservable factors, we will treat it as a "residual" explanation and be more inclined to accept it only if the first one fails to account for the set of observations[62].

Finally, we also observed that increasing the level of temporal aggregation can indeed introduce GARP violations to datasets that contain no high-frequency inconsistencies. Whether this last result is simply a mathematical curiosity or a likely circumstance of actual data is an empirical matter, which will also be considered next.

[61]Problem 2 implies a pair of nonintersecting annual budget lines, whereas in problem 3 those lines must still intersect at the lower frequency; unless the studied sample covers only a couple of years, both can occur for different pairwise comparisons of periods.

[62]Occam's razor considerations could also be used to support this decision, evidently.

3.3 Experiments on GARP and time aggregation

In the previous section we demonstrated how high-frequency GARP violations can go undetected in low frequency datasets. We also observed that the opposite case is possible: one can find GARP violations in, say, annual data even if quarterly or semesterly choices pass the GARP test. Now we will study those same aspects using simulation exercises previously proposed in the literature to evaluate the performance of the GARP test. We will start with a slightly changed version of the exercises proposed in Fleissig and Whitney (2003): to simulate GARP-consistent data with and without measurement errors and verify to what extent time aggregation affects the test's ability to point out inconsistencies. Next, we will adopt Bronars' (1987) standard simulation approach to check the power of the GARP test in actual datasets of consumption expenditures, at multiple frequencies. What can be considered shortcomings of Bronars' method, as well as possible improvements, will be discussed later in section 3.4.

3.3.1 Temporal aggregation and GARP consistency of Cobb-Douglas Data

Fleissig and Whitney (2003) studied, among other things, the sensitivity of GARP and weak separability tests to the occurrence of measurement error. Part of their simulation exercises involved generating Cobb-Douglas demand data (which is GARP-consistent and weakly separable, by construction), adding random shocks of some expected magnitudes to all series and checking how often the tests reject their respective null hypotheses. The data generating method for five goods and 40 observations had essentially four steps:

1st) Randomly draw a series of 40 observations for total expenditures from the uniform distribution **U**[10,000, 12,000];

2nd) Randomly draw five series of 40 observations for the prices of each good from another uniform distribution **U**[98, 100];

3rd) Letting alpha≡[0.60; 0.25; 0.10; 0.04; 0.01] be the vector of coefficients on a Cobb-Douglas utility function and given the series on total expenditure and prices for each good, calculate x_{it}, the Marshallian demand for good i at period t, without measurement errors.

4th) Introduce random measurement errors to the demands for the last four goods according to the formula $x_{it}^{error} = x_{it} * e_{iK}$, where $e_{iK} \sim \mathbf{U}[1-K,1+K]$ and K is the maximum magnitude of the "shocks" at each time, $K \in \{0.01;0.05;0.10;0.20\}$ for 1%, 5%, ..., etc.

5th) To calculate the demand for good 1, divide residual income by its price, at each period.

Fleissig and Whitney (2003) actually repeated those steps with three different price distributions and also using an alternative vector of preference parameters – that add up to one, obviously. In all cases, the procedure successfully generated data with abundant budget hyperplane intersections. The tests were applied on 2000 simulated datasets for each particular specification (combination of price distribution and preferences parameters).

Following essentially the same steps and using two of their vectors of preference parameters, we study once more the performance of the GARP test in the presence of measurement error, but we also consider how time aggregation affects the results. Annual series of 37 observations for each of the five goods were calculated with the aggregation (averaging) of 148-observation "quarterly" samples – simulated as described above; the lengths of simulated series match the sample sizes of our actual consumption data (1964-2000, to be studied later) because we want to make comparisons between difference pieces of evidence that will be collected throughout this section; the GARP test being nonstochastic, the chance of finding a GARP violation and rejecting the consistency of a simulated dataset that contains measurement errors increases with the sample size. In other words, longer/shorter simulated series would be more/less likely to

contain violations than actual figures simply due to their sample size, making results not comparable.

Table 3.4 shows how often Cobb-Douglas data constructed with and without errors did not contain a single GARP violation, at quarterly (originally simulated figures) or annual frequencies (time-aggregated data), out of 2000 trials each time; each half of the table shows results for data generated with particular sets of preference coefficients also adopted by Fleissig and Whitney (2003), as a way to check for the robustness of results. Notice first that 100% of the annual datasets passed the GARP test; recall that in section 3.2 we verified the possibility of finding violations in low-frequency data calculated from GARP-consistent high-frequency series – even if budget lines intersected at all frequencies. We can deduce from this first piece of evidence that GARP violations at annual but not at quarterly or monthly series will be unlikely in datasets of actual consumption expenditures that can be rationalized by Cobb-Douglas preferences at high frequencies.

As we allow for measurement errors of different magnitudes, the same pattern is revealed in both halves of the table: regardless of preference parameters in the utility function, the test detects inconsistencies much less often with averaged (annual) data than otherwise; the proportion of datasets passing GARP obviously decreases with the introduction of possibly larger measurement errors, but the decay is much less severe with low-frequency data. Allowing for up to 5% measurement error makes the occurrence of GARP violation at least twice as likely in the originally simulated (quarterly) data than with time-aggregated (annual) figures. As discussed before, results for datasets a) and b) in a given Cobb-Douglas specification are not exactly comparable to each other because the likelihood of GARP violations is expected to increase with sample size; for that reason, we also generated 4 times longer samples of quarterly data and averaged them up to generate "annual" datasets with the same sample size of quarterly ones. Comparing results from datasets a) and c), we conclude that time aggregation indeed reduces the test's ability to detect inconsistencies even if one controls for the sample-size problem.

Finally, we acknowledge that not all of the time-aggregation effects analytically discussed in the previous section can be associated with our findings from this first exercise. The less frequent detection of GARP violations in time-aggregated data can be partially attributed to the relative price smoothing effect, as high-frequency relative prices are expected to have a larger variance than their annual counterparts, obviously; however, the data simulating method used here involves drawing income values from a uniform distribution, which contrasts sharply to our discussion of real expenditures shifting effects and their likely relevance in datasets where real expenditures tend to grow over time. This issue will be addressed in the next simulation approach, which uses information from actual consumption datasets.

3.3.2 Applying Bronars'(1987) approach to study the power of the test

Bronars' simulation approach was proposed to address particularly the aspect of temporal aggregation that was not incorporated in our first exercise: the expansion of budget sets over time due to recurrent positive income shifts in actual data. He conditioned the acceptance of the GARP test results to a more careful examination of the data; as observed choices can only be considered "revealed preferred" to feasible alternatives, no GARP violation can occur if at each period the consumer can purchase all affordable bundles of previous moments.

The method is rather appealing; it essentially involves simulating random budget share allocations along the observed budget hyperplanes, which can be trivially deduced from datasets of actual prices and quantities. By doing so, it incorporates Becker's (1962) notion of irrational (random) behavior as the true alternative hypothesis and evaluates how often the test is able to reject its null hypothesis of GARP consistency in randomly simulated data. If the series with actual choices pass GARP and the simulated ones fail to do so reasonably often, there is evidence against the view that the nonrejection of the actual choices was due to the absence of budget intersections; the

consumer actually seems to behave as an utility maximizer with a stable ordering of preferences for alternative combinations of goods. The power of the test against the alternative hypothesis in a specific dataset is measured as the percentage of times that GARP is rejected over several simulations of random data, all constituting bundles on the actual budget hyperplanes[63].

Two algorithms used by Bronars to generate random points along observed budget hyperplanes are of particular interest here. For n commodities and t periods, the first algorithm involves drawing random variables $Z_{1t},\ldots, Z_{nt}$ from a uniform distribution so that random budget shares S_{it} allocated to the purchase of each good i are calculated as below[64]:

$$S_{it\,(1)} = Z_{it} \Big/ \sum_{j=1}^{n} Z_{jt} \qquad \textbf{(3.19)}$$

The sum of S_{it} for all n goods is 1, meaning that even an "irrational" consumer is expected to exhaust the income available at each time by picking random points along – and never below – the observed budget lines. In other words, his irrationality actually involves spending all his income randomly, on the available goods. Having simulated the series on budget shares, real expenditures on each good are calculated by multiplying those shares by total expenditures, and the resulting figures are divided by actual prices (price indices) of the corresponding commodities. The final numbers constitute proxies of real quantities demanded by an irrational consumer.

One can expect that testing GARP-consistency of purely random allocations of resources leads mostly to rejections of the hypothesis of utility maximization but that does not necessarily happen, especially if the actual budget lines shift significantly over

[63] Also central to his discussion was the fact that the nature and the frequency of budget line intersections would differ significantly if the researcher adopted aggregate or per-capita consumption figures as the baseline for those simulations; his empirical finding – consistent with such explanation – was that the power of the test was much higher in per capita datasets than with aggregate series.

[64] The algorithms referred here as "first" and "second" are actually Bronars'(1987) second and third, as he also considered a simpler one, with random numbers being drawn from the uniform distribution.

time; if budget lines do not intersect in the actual dataset, they will also not intersect within randomly simulated series, and GARP violations will never occur in both cases.

Bronars acknowledged that the algorithm above implies an expected budget share of 1/n for each commodity, and also that actual purchases of some goods typically represent larger shares of total expenditures than others. For reasons that will be discussed subsequently, he proposed another method to generate random figures but this time making the expected budget shares allocated to a specific commodity the same in actual and simulated data:

$$S_{it\,(2)} = K_i Z_i \Big/ \sum_{j=1}^{n} K_j Z_j \qquad \textbf{(3.20)}$$

where K_i is the mean budget share of good i in the actual data across all years. Generated like this, the simulated budget shares for each good will randomly fluctuate around their historical (sample) average.

Bronars justified the use of both algorithms above stating that they make unlikely the picking of bundles near intercepts by the irrational consumer; the expected budget share simulated for any good with the first algorithm is 1/n (n being the number of goods), whereas with the second method the expected budget share is simply the historical average share allocated for that good. In any case, however, the algorithms are unlikely to simulate budget allocations that result in most of the income being spent, by chance, on a single good. He claims that such a fact matters because budget intersections in post-war U.S. aggregate consumption data occurred mostly near intercepts; consequently, the use of his algorithms would prevent an overestimation of the power of the test in his dataset.

Much in the same way Bronars applied those techniques to compare the power of the test over aggregate and per capita data, we now reconsider this aspect of the test using per capita data at different frequencies. Table 3.5 shows the power of the test over four distinct datasets, two of them at each frequency; the robustness of conclusions is

checked with respect to a possibly relevant aspect: the exclusion of leisure from the set of consumption subcategories. Q1 and A1 refer respectively to quarterly and annual datasets on all subcategories of durables, nondurables, services and leisure – data sources and methods precisely the same ones discussed in chapter 1; the last two datasets at each frequency exclude leisure (Q2, A2).

The most important reason to check the power of the test on datasets with and without the leisure figures is the fact that the frequency of budget intersections could be significantly affected[65]. Suppose, for example, that an exogenous positive shock to the "price" of leisure – the opportunity cost of hours not worked – is not followed by significant responses on the number of hours worked, due to labor contract rigidities; the budget hyperplanes for periods before and after the shock may become further distant from each other, as only an income effect is observed. On the other hand, large fluctuations of the price of leisure can also increase the frequency of budget intersections over time, using the same logic discussed above, provided that they are not always positive or negative. The issue can be empirically investigated with simple computations of how often budget hyperplanes do intersect within both datasets, as we also report in table 3.5.

As we inspect first the results for quarterly data (Q1 and Q2 in table 3.5), the power of the GARP test against random behavior is – strikingly – the highest possible. Not even once (out of 2000 simulations) did the test fail to reject GARP over random data, regardless of data simulation methods or inclusion/exclusion of leisure. As for the number of times individual budget hyperplanes (for a given quarter) intersected with the ones for different periods, the frequency is also surprisingly high with and without the inclusion of leisure, but more so with it. Even without the inclusion of leisure, however, the median number of times a given budget set intersects with others is very large,

[65]We acknowledge that the exclusion of any good in the context of the GARP test is equivalent to assuming *a priori* that all other goods are weakly separable from it in the utility function; further, that such hypothesis was empirically rejected at the quarterly frequency in chapter 1, with the very same data adopted here. However, we also performed our analysis over datasets without leisure because we are well aware of the criticism regarding the usual estimates of leisure prices and quantities. We will return to those issues in chapter 4.

around 60% of the possible number of times. Those results[66] seem to corroborate the high power of the test over datasets with often intersecting hyperplanes, but they can also be a reason for concern; we will return to this possibility soon.

The results with annual data indicate that the power of the GARP test is indeed reduced within low-frequency datasets. Once more, the inclusion of leisure tended to raise the power of the test against the alternative hypotheses: the rejection rates with A1 are larger than with A2 with both data simulation methods; such higher power is consistent with the fact that any given annual budget hyperplane is more likely to intersect with others in datasets that include leisure – the minimum number of budget intersections jumps from 8 out of 42 possible intersections to 29 out of 37, as leisure is included.

As one compares the rejection rates for each of Bronars' alternative hypotheses or irrationality models, the test is revealed to be much weaker against $S_{it}(2)$; focusing solely on the first algorithm would lead one to believe that the test is very powerful with annual datasets, but it is so only against that very particular model of irrational behavior. This is just what can be considered a limitation of Bronars' approach, as we discuss next.

[66]All tests were repeated over datasets that assumed two different expectation schemes in the calculation of user costs of durable goods, following FHS (again, details in Appendix); since results were not sensitive to such assumptions, the only figures reported in table 3.5 are the ones for the case of perfect foresight, for simplicity. A complete table with rejection rates of all tested datasets, using the simulation methods discussed in this chapter, is available upon request.

3.4 Weakness and extensions to Bronars' approach

The core of Bronars' power measure for the GARP test is the definition of an alternative assumption, hereafter also referred to as the "irrationality" model. Data are generated with criteria other than the maximization of well-behaved utility functions, so that one can evaluate how often the test is able to distinguish between actual and simulated choices made along a set of observed budget lines. That is also the source of its most notorious shortcoming: any conclusion from the application of such method is conditional on the particulars of the budget-share simulating algorithm. Unless the test is shown to have high power in a dataset regardless of the adoption of multiple irrationality models – like the case of quarterly data in the previous section – it is hard to tell how strong is the evidence of utility-maximizing behavior coming from the GARP-consistency of a dataset, as in the case of annual data before. If the test seems to have high power against the alternative hypothesis X but not against Y, what is the conclusion?

Rather than trying to find a general answer to that question, in this section we show that some simulation algorithms can be considered inadequate under certain circumstances. First, however, a couple of alternative algorithms proposed in the literature as extensions to Bronars' approach will be considered; next, we combine the two simulation exercises implemented before to show how the algorithms can fail their purpose of generating "irrational" data. Then we will suggest the use of the time series techniques adopted in chapter 2 to obtain information on the evolution of actual budget shares. It will be argued that obtaining such information is an important step preceding the interpretation of results from Bronars' approach. We conclude the section proposing a new simulation algorithm, which may have advantages in replicating relevant characteristics of actual data.

3.4.1 GARP consistency of simulated random data: discussion and evidence

Recall once more Bronars' second algorithm $S_{it}(2)$, in which consumption choices are still random but the mean budget share of simulated data equals the historical average figure for some data sample. Unless actual series on budget shares behave as stationary processes in finite samples[67], the allocation of resources on consumption subcategories are expected to change over time, not necessarily returning to an overall average value. Bronars' algorithm based on the mean budget shares seems to incorporate, in that case, a somewhat irrelevant characteristic of the original choices into simulated figures. What is more concerning, however, is that since the series on simulated budget shares are built to fluctuate around a fixed value, they can end up revealing rational behavior, rather than otherwise: the corresponding simulated demands can approximate those for weakly separable goods, derived from a Cobb-Douglas utility function with well-behaved (uncorrelated) errors[68]. GARP violations would even tend to disappear, if the simulated choices contained only relatively small budget share fluctuations around the overall averages. Of course, that possibility is unlikely if the studied dataset is long and, consequently, the chance of drawing only random figures that differ little from the overall expected value is very small. In any case, the weakness of Bronars' second simulating algorithm is evident: it can fail to generate irrational behavior, especially when applied to small samples. The GARP test may (correctly) not reject the supposedly irrational choices, leading the researcher to conclude that the power of the test is low in the studied dataset, which may not be true.

The possibility of generating random GARP-consistent choices and, for that reason, underestimating the power of the test in a dataset is indeed much more likely if the researcher adopts two recently proposed simulation algorithms, as we discuss next.

---

[67]Hamilton (1990) defines a time series process as stationary if neither its mean nor its autocovariance depend on the date; intuitively, such process tends to return to an overall average value after transitory fluctuations fade away. We will return to this issue in a coming subsection.

[68]It is well-know that Cobb-Douglas demand data imply fixed budget share allocations for each good.

Burton (1994) applied and extended Bronars' approach in the analysis of British meat and fish consumption data. He actually used Bronars' data simulation methods, besides one of his own: he generated uniformly distributed budget share allocations along the ranges of historically observed choices, i.e., in bounded regions of the consumer's hyperplane. Let W_{it} be a random variable drawn from the uniform distribution $U[min_i, max_i]$, where min_i and max_i are the extreme budget share historically allocated to some good i, respectively; then:

$$S_{it}(3) = W_{it} \quad \text{for i=1,…, n-1} \tag{3.21}$$

$$S_{nt}(3) = 1 - \sum_{i=1}^{n-1} W_{it} \tag{3.22}$$

Notice that whenever the sum of simulated budget shares for the first (n-1) goods exceeds one, or if the residual budget share for the n^{th} good is outside the interval of actually observed choices, all numbers are dropped and new draws are made.

This algorithm was proposed as an improvement of Bronars's methods, for it avoids simulated budget allocations that are highly untypical of actual data. The same way Bronars wanted to avoid overestimating the power of the test by making the simulation of "extreme" budget share allocations less likely, Burton (1994) felt that the bar could be raised some more: the test would only be considered strong enough if it could actually detect GARP inconsistencies in the range of shares allocations observed over the sample. Before we explore the likely caveat of this method, let us briefly present the second alternative to Bronars' algorithms that will eventually have the same problem.

Cox (1997) proposed another alternative hypothesis to introduce "irrationality" through randomness but preserving characteristics of real data. He proposed the reassignment of observed budget share figures randomly throughout a sample; actual budget shares observed at some period t would be used as the "simulated" budget share for period t+r. These "random" budget allocations are consistent with actual observed

choices to some degree, but they are clearly independent of prices, by construction. Hereafter we will refer to Cox's irrationality model as the fourth (model 4), and the corresponding simulated budget shares as $S_{it}(4)$.

To see how both alternative algorithms are problematic under fairly common circumstances, consider what the estimated power of the test would indicate if the true budget shares allocated to the consumption of some/all of the goods were to fluctuate little around fixed values throughout the sample. An interesting way to do just that is to combine the two simulation exercises implemented so far. We will generate a large number of Cobb-Douglas datasets, following precisely the procedures in our first simulation exercise, and submit them to Bronars's approach using all simulating algorithms. Notice that the interpretation of results in this case is slightly changed, compared to the standard application of Bronars' approach. Since the datasets are actually constructed in a way that guarantees abundant budget intersections, the % number of times datasets are rejected indicates not the power of the test in those series – which is high, by construction – but the ability of an algorithm to generate data that are not consistent with GARP. The results of this third simulation exercise are presented in table 3.6, which reports rejection rates for simulated data based on demand figures generated from two Cobb-Douglas specifications.

Both Burton's and Cox's algorithms simulate budget shares that are bounded inside the range of actual observations; when the actual data are generated without measurement errors (K=0%), actual and simulated figures in those cases are identical and the test never rejects the supposedly random figures. If the researcher is unaware of the true data generating process and applies those two algorithms to evaluate the power of the test, he will conclude that it is very low. But if one uses Bronars methods instead, the conclusion is quite the opposite. The data simulated with Bronars' algorithms fail to pass GARP at least 95% of the time because budget intersections occur frequently enough and the test is truly powerful in that dataset. Also interesting is the fact that the use of Bronars' original method to evaluate the power of the test do not seem to be

sensitive to the magnitude of measurement errors, with rejection rates varying little with the introduction of larger errors.

Finally, we acknowledge that Cox (1997) himself observed that the GARP test would likely fail to reject his random (simulated) data in the sort of circumstance discussed above, leading to biased conclusions about the power of the test. Concerns of this nature motivated our proposal of a new data simulation method, to be presented later; an algorithm that would not tend to underestimate or overestimate the power of the tests under different circumstances.[69,70]

3.4.2 The evolution of budget shares in actual datasets

Our discussion in the previous subsection indicates that prior knowledge about the evolution of budget shares in actual consumption data may help in one's interpretation of findings from the applications of Bronars' approach. We now use Rossana and Seater's method (discussed and adopted also in chapter 2) to check whether budget share allocations underlying U.S. consumption expenditures behave as stationary processes in our samples. If they do, quarterly and/or annual figures on budget shares can be expected to return to their long run averages after eventual temporary shocks, with two interesting implications: first, the demands for those goods may be reasonably well characterized as Cobb-Douglas ones, especially if deviations of budget shares from their long run averages are relatively small; consequently, we can suspect that the use of Burton's and Cox's algorithms to further evaluate the power of the test in those datasets is somewhat inadequate, for they may underestimate the power of the test in such

[69]Also worth mentioning, Aizcorbe (1991) proposed an alternative to Bronars (1987) simulation approach, suggesting a lower bound for the power of GARP. However, her method provides little information if the dataset involves a large number of goods or a small number of observations, as in her replication of Bronars analysis over 9 consumption categories. The set of consumption categories studied here is even larger, including 14 goods and services besides leisure.

[70]See also Gross (1995) for a GARP-based method of hypothesis testing particularly well suited to the investigation of commonality of preferences among consumers.

circumstances, as discussed before. Finally, a last implication is that Bronars second algorithm would not be generating a somewhat "unnatural" behavior in simulated figures by constructing them to fluctuate around the sample average.

The best-fitting specifications[71] for quarterly and annual series on budget shares are reported in table 3.7. The hypotheses that each series contained a unit root were never rejected at the 5% significance level with both quarterly and annual data, regardless of the ADF test specification (including a constant or both constant and trend terms). As one compares the overall results from the two columns, almost all elected ARIMA specifications had longer and more complex lag-structures with quarterly data than with annual figures. Also, simple model specifications such as IMA(1,1) or even the random walk model generated white-noise residuals much more often with annual data than with quarterly series. Those results are pretty much consistent with our findings discussed in chapter 2, as well as with Rossana and Seater's (1995) original conclusion that time aggregation tends to cause a significant loss of information regarding the time series processes driving the variables.

Before proceeding, a word of caution is still due. It is known that because series like the ones on budget shares are bounded (between 0 and 1), they are, technically, stationary; however, our results reported in table 3.7 indicate that those series indeed behave as nonstationary processes on finite samples. Even though model specifications were only considered good candidates if they generated serially uncorrelated residuals, rather than difference-stationary the unknown data generating process could actually be trend-stationary, with the presence of structural breaks over the sample making identification a much more complex task[72]. In any case, the undisputable piece of evidence provided by this short exercise of time series analysis is that the series on budget shares do not seem to fluctuate around their long-run averages. That fact, itself, permits our conclusion that the Bronars' second algorithm [$S_{it}(2)$] incorporates into the simulated series on budget shares a characteristic that is not likely present in actual

[71]See chapter 2 for details on Rossana and Seater's procedures.

numbers. Rather than simply irrelevant, such a feature can lead to the underestimation of the power of the test in relatively short samples, as discussed before. That seems to be a reasonable explanation for the results reported previously on table 3.5, according to which the use of Bronars' second algorithm suggests a much lower power of the test in annual datasets (with shorter samples than the quarterly ones) than one can observe with his first method.

3.4.3 A new simulation algorithm: incorporating consumption trends

In this subsection we introduce a new simulation algorithm and check the robustness of the results reported in table 3.5 by applying all simulation methods discussed so far to our quarterly and annual datasets.

Regardless of data frequencies, we observed that quarterly and annual budget shares do not seem to fluctuate around overall historical values. As those figures are bounded inside the interval [0,1] by construction, it may be informative to further consider how large a range of values the shares of individual consumption categories to assume throughout our sample. Table 3.8 reports the minimum and maximum figures on the % share of consumption expenditures for each good (annual data), including or not leisure estimates as part of expenditures. While many consumption subcategories seem to wander inside relatively narrow ranges of budget share allocations (ND4, ND5, S1, S2, S3), other such as Food (ND1) and Medical Care (S4) seem to have changed a lot over those years.

We now propose a straightforward adjustment to Bronars' second algorithm [$S_{it}(2)$] that permits incorporating all those patterns into the simulated series, regardless of the specifics of the unknown data generating process for each of the 15 goods. It simply involves replacing the historical average budget shares across all years (K_i) in

[72]Enders' (1995) chapter 4 summarizes what is actually learned and how the researcher should proceed as the unit root hypothesis is not rejected at the first stages of model identification.

expression (3.20) with moving averages (L_i) of observed budget shares across some range of periods around each time t; the expression would then become:

$$S_{it\,(5)} = L_{it} \cdot Z_{it} \Big/ \sum_{j=1}^{n} L_{jt} \cdot Z_{jt} \qquad \textbf{(3.23)}$$

where L_{it} corresponds to the mean budget share of good i in the actual data over the arbitrary interval $[t-\tau, t+\tau]$. Generating random fluctuations around a moving average of actual figures can accomplish two goals: it maintain the randomness of choices without systematically imposing any atypical evolution of budget shares throughout the sample.

A quick look at actual and simulated budget shares from Bronars' second algorithm [$S_{it}(2)$] and our own [$S_{it}(5)$] can further highlight our contribution. As shown in figure 3.10, U.S. consumers have increased significantly the share of their resources allocated to the consumption of medical services over the period 1964-2000; the smooth line moving from 1.8 to 5.1% of (annual) total consumption expenditures represents the evolution of budget shares actually allocated to the purchase of those services over time. To illustrate how different the simulated random budget shares based on those two algorithms may be, we drew a single set of random numbers from a uniform distribution and applied them along with actual figures to each algorithm. Initially, one may think that there is not significant difference between data simulated with those algorithms, as they seem to vary in a similar fashion around the actual figures.

A closer look reveals that the actual budget share of medical services increased smoothly and almost monotonically over the period 1964-1993, remaining stable thereafter[73]; with the same series of random shocks, however, Bronars' use of a historical average budget share tended to overestimate the simulated figures for the first years and underestimate them at the end of the sample, relative to the new method. The

[73]The choice of figures on medical services is specially interesting because even though the presence of a unit root was not rejected in this series, it could also be described as a trend-stationary process with a structural change in the deterministic trend.

two simulated series are only similar at the middle of the sample, as local and overall averages tended to coincide over those years (early 80's).

The new algorithm [$S_{it}(5)$] permits investigating the power of GARP test against a slightly changed alternative hypothesis: the consumer still chooses random points along their budget hyperplanes, but in such a way that the simulated budget shares tend to evolve similarly to the actual figures over time. Rather than using the overall average as the parameter from which simulated budget shares would randomly deviate at each period, we arbitrarily picked the mean budget share of 5 periods over the local interval [t-2,t+2]. Four observations were lost, but we gained in the sense that these simulated data do not deviate systematically from the actual evolution of budget shares over the sample. Thus, the new method preserves eventual changes in budget shares due to long-run permanent changes in relative prices over those periods.

Table 3.9 reports estimates of the power of the test using all simulation algorithms discussed so far (including once more the number of budget intersections observed in each dataset). The use of all 5 methods generally confirmed, to some degree, the overall conclusions previously drawn from table 3.5: the power of the test is indeed very high within quarterly datasets, regardless of leisure inclusion or simulation methods; also, the tests seem to be substantially weaker at the annual frequency against the majority of alternative hypotheses.

Randomly simulated quarterly datasets were rejected quite often, at least 96% of the times. As with Bronars' simulation methods [$S_{it}(1)$, $S_{it}(2)$], all 2000 quarterly datasets were rejected with the new method [$S_{it}(5)$], but not when the other two alternative irrationality models were used [$S_{it}(3)$, $S_{it}(4)$].

Nevertheless, the discrepancy of findings from the three alternative irrationality models is more striking with annual data. First, notice that the power of the test is extremely low against the hypotheses suggested in Burton (1994) and in Cox (1997), with rejection rates for annual data below 30%. Second, the results from the use of $S_{it}(3)$ and $S_{it}(4)$ were also atypical with respect to the inclusion of leisure, which only in those cases tended to reduce the power of the test. Finally, the new data simulation method

$S_{it}(5)$ led to rejection rates always lower than the numbers from Bronars' first algorithm, $S_{it}(1)$, but higher than the ones for his second method, $S_{it}(2)$.

The first two issues raised in the last paragraph can be addressed together, as they are pieces of the same puzzle: the test detected no GARP violations 70% or more of the time that random annual data were generated according to algorithms $S_{it}(3)$ or $S_{it}(4)$; however, budget intersections still occurred quite often at that frequency – at least 8 times but typically more than 25 times. How can that be? The answer is that randomness does not necessarily imply irrationality. The common characteristic of those two simulation methods is that they are strictly limited by actually observed choices. Regardless if one is simply randomly rearranging budget shares along the sample period [$S_{it}(4)$] or drawing random points from ranges of extreme choices [$S_{it}(3)$], the simulated budget shares in those cases may never assume values substantially different than the actual (rational) ones. Since actual budget shares for many goods lie along relatively narrow ranges of values, the chances of simulating data that are significantly different than the (actual) rational allocations of resources with both algorithms are simply too small. The introduction of leisure seems to expose further the caveats of simulations methods $S_{it}(3)$ and $S_{it}(4)$; as depicted in table 3.8, the inclusion of leisure involves implicitly assuming that an outstandingly larger budget share is allocated to a single good at all times. It necessarily makes narrower the observed ranges of budget shares of other goods. Note, for example, that the share of expenditures on food (ND1) ranges from 14.7 to 26.0% of total expenditures if leisure is excluded, and only from 4.9 to 5.4 with it. Therefore, a plausible explanation for that apparent puzzle is that those simulation methods have merely failed to generate irrational data.

As for the last issue, the estimated power of the test may have been so lower with the second algorithm than with the first because it actually generates data that are, on average, consistent with GARP. The new simulation method apparently has produced figures that are neither far too "extreme" as in Bronars' $S_{it}(1)$ nor so likely close in small samples to those of a truly rational consumer, as in all other ones. $S_{it}(5)$ seems to be a reasonable alternative because it preserves the "irrationality" of random budget shares

that are not bounded inside narrow intervals – a deficiency of $S_{it}(3)$ and $S_{it}(4)$ under certain circumstances – but also eventual long-run trends of actual consumption data that are not incorporated with Bronars' methods[74].

3.5 Final remarks and conclusions

In this chapter we have shown from both an analytical and an empirical perspective that the GARP test can lead to biased conclusions on the existence of well-behaved utility functions rationalizing low-frequency datasets. First, we formally described the manner in which GARP violations observed within quarterly datasets may be undetectable over annual series; two factors commonly discussed in this literature – relative price changes and real income shifts – were shown to be particularly relevant to explaining how temporal aggregation affects GARP because they could reduce the frequency of budget intersections in annual series, relatively to quarterly ones. However, it was also demonstrated that temporal aggregation can eliminate high-frequency GARP violations even when time-reaggregated data still contained budget intersections, and the power of the test was not to blame. Therefore, the discrepancy among results at different frequencies indeed demanded further empirical analysis on the power of the GARP test.

We found that the power of Varian's nonparametric test against alternative hypotheses of random behavior tends to be substantially lower as one adopts annual rather than quarterly per capita consumption data. In other words, the test was much more likely to commit a type II error and accept the utility maximizing hypothesis with annual random data than with quarterly figures. This main conclusion is consistent with Rossana and Seater's (1995) perception that temporal aggregation tends to cause

[74]Despite the evidences gathered here, extensive exercises of data simulation are necessary before the superiority of any algorithm is formally established; that is, apparently, a promising venue for future research.

significant loss of information about the true processes driving economic variables (confirmed by our analysis of the revised NIPA data in chapter 2).

The low power of the test within annual datasets raises concerns and suggests a revision of previous findings and maintained assumptions in empirical macroeconomic studies. First, it provides a simpler explanation than Swofford and Whitney's (1987, 1988) for the nonrejection of GARP-consistency on annual datasets including broad monetary assets (Money-in-Utility Function models); they argued, instead, that the existence of short-run costs to adjust illiquid asset holdings could explain the occurrence of GARP violations within quarterly but not annual datasets. The new explanation based on the power of the GARP test has the advantage of not relying on unobservable factors (adjustment costs). Furthermore, there seems to be no reason to believe that researchers are better off adopting annual rather than quarterly data to study such models, as for example in the more recent contribution of Holman (1998).

As a second implication, our (non)rejection of particularly important separability structures in chapter 1 can be considered a strong result, for the reasons we repeat below. Empirical Consumption-Asset Pricing models have, with very few exceptions, assumed the weak separability of Nondurables and Services from all other consumption subcategories. In chapter 1 we saw that such a hypothesis was rejected at the quarterly frequency but not with annual data. An alternative preference structure passed both necessary and sufficient conditions for weak separability at both quarterly and annual frequencies: the one including Nondurables, Services and Leisure in the representative utility function. Since we found that the power of the test was extremely high within quarterly datasets, both the rejection of the commonly maintained assumption and the nonrejection of its alternative at the quarterly frequency can be considered very strong results. On the other hand, caution is recommended with respect to other results in that same chapter. A large variety of separability structures passed GARP consistency and weak separability tests only at the annual frequency. Those findings should be interpreted as further evidence that the test is more likely to commit type II errors within

low-frequency datasets, rather than as indication that datasets at the annual frequency are preferable.

Over the last two sections we made use of multiple simulation exercises and studied the test's ability to detect inconsistencies in actual and constructed datasets. We found that Bronars' (1987) framework is susceptible to generating biased estimates of the power of the test because some algorithms used to generate random data are likely to produce GARP-consistent figures, under certain circumstances. A brief investigation on the time series properties of actual budget shares permitted our suggestion of a new data simulation method, which seems to have advantages over existing algorithms. Extensive use of Monte Carlo experiments seems to be a promising way to establish the superiority of any of those algorithms, but that is beyond the scope of the current research project.

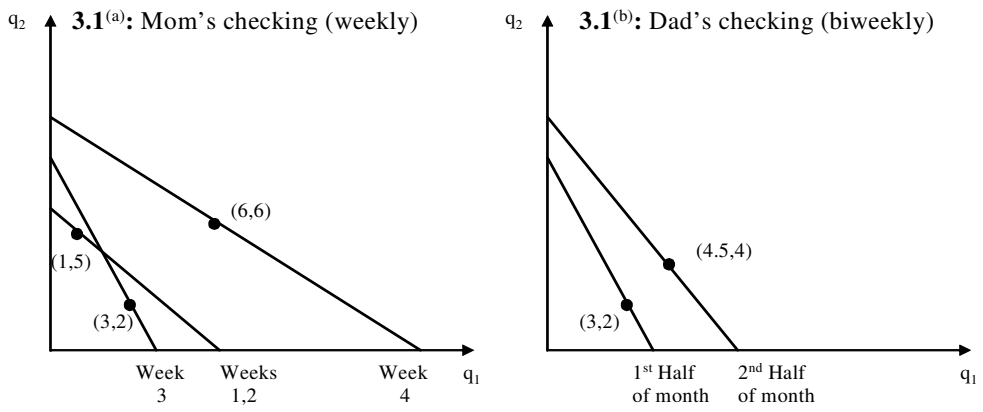

Figure 3.1: Weekly consumption of a teenager

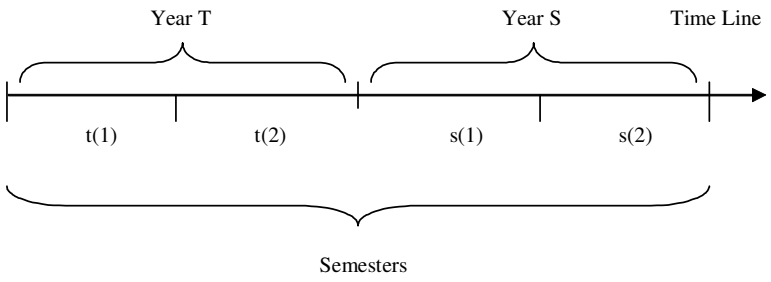

Diagram 3.2: The notation of periods

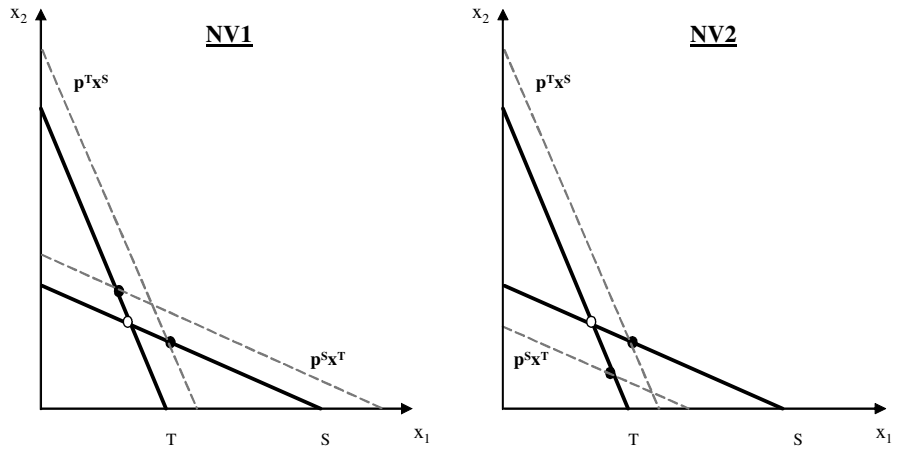

Figure 3.3: GARP-consistent choices along intersecting budget lines

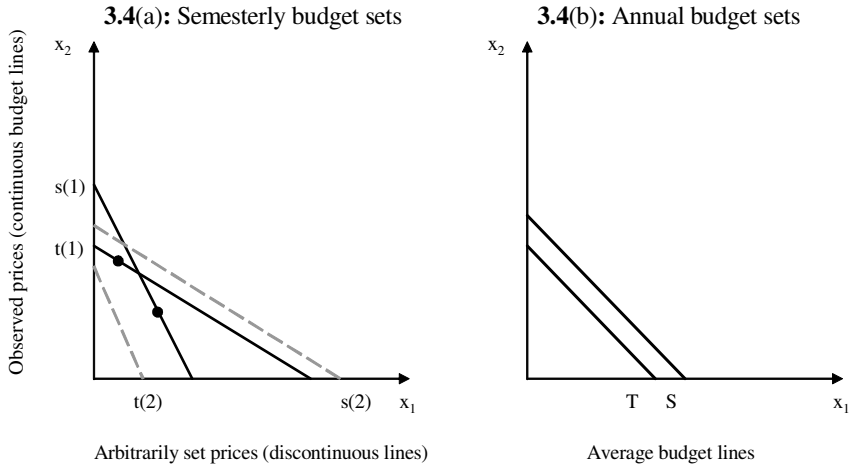

Figure 3.4: Illustrating the RPS effect of temporal aggregation

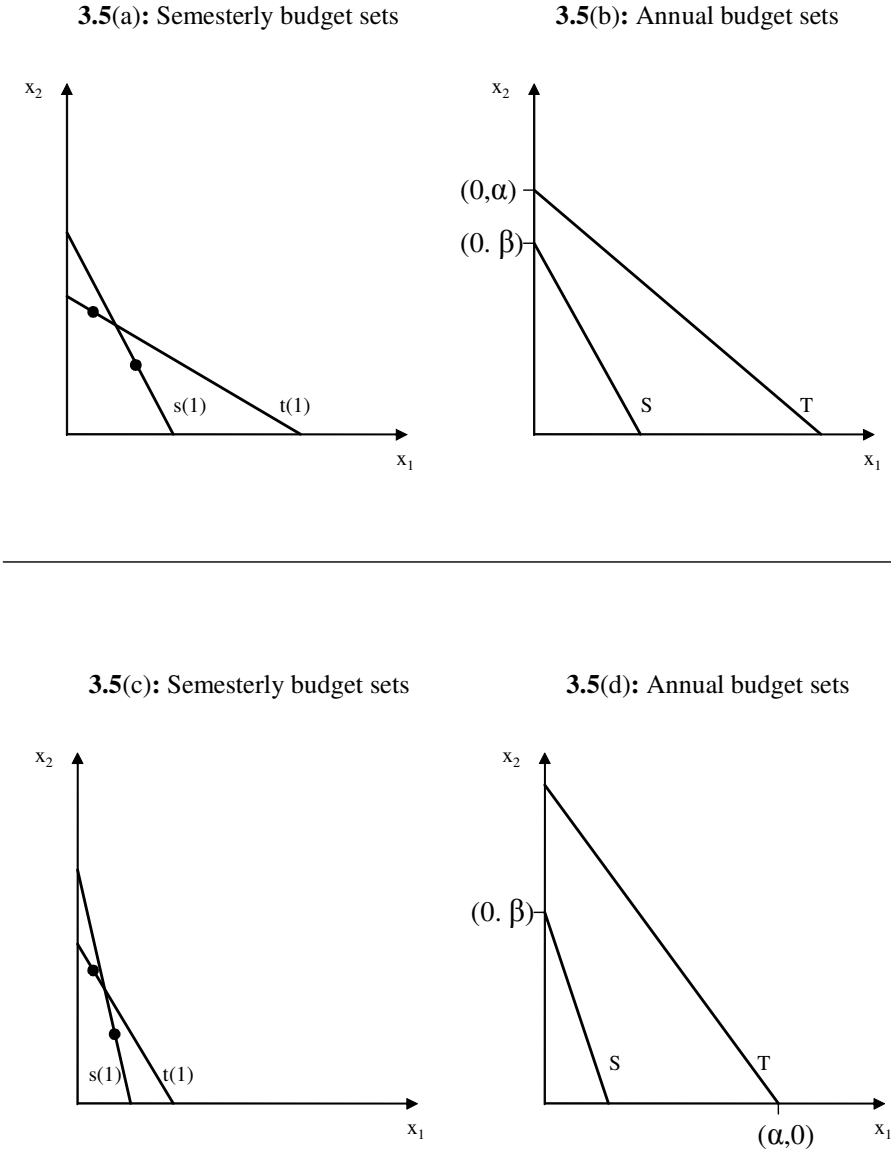

Figure 3.5: Two illustrations of the RES effect

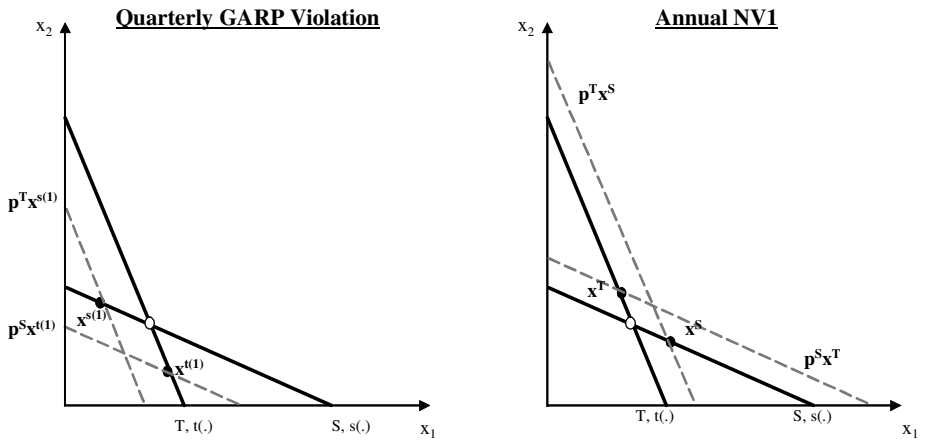

Figure 3.6: GARP violation with quarterly data, but not with annual figures

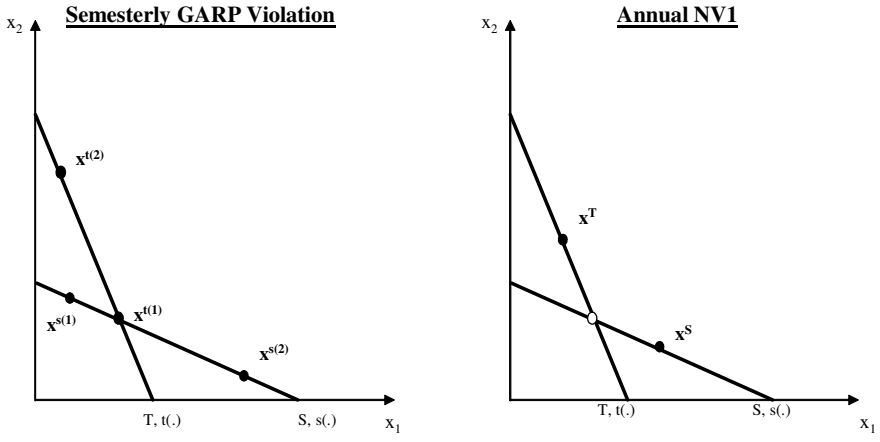

Figure 3.7: A second example of GARP violation with semesterly data (only)

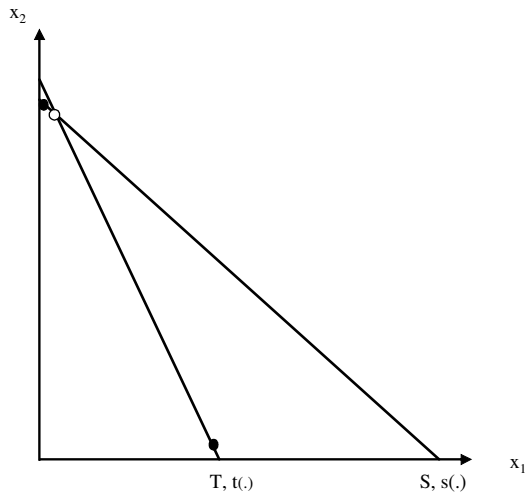

Figure 3.8: "Extreme" GARP violation

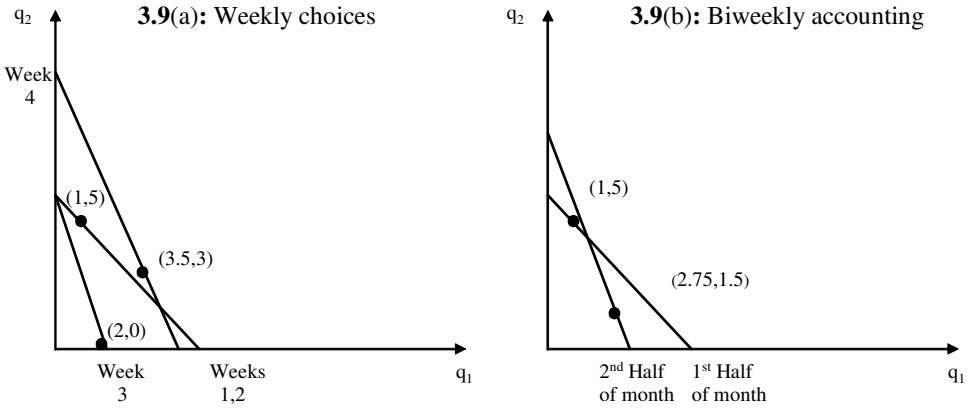

Figure 3.9: GARP violation as a result of temporal aggregation

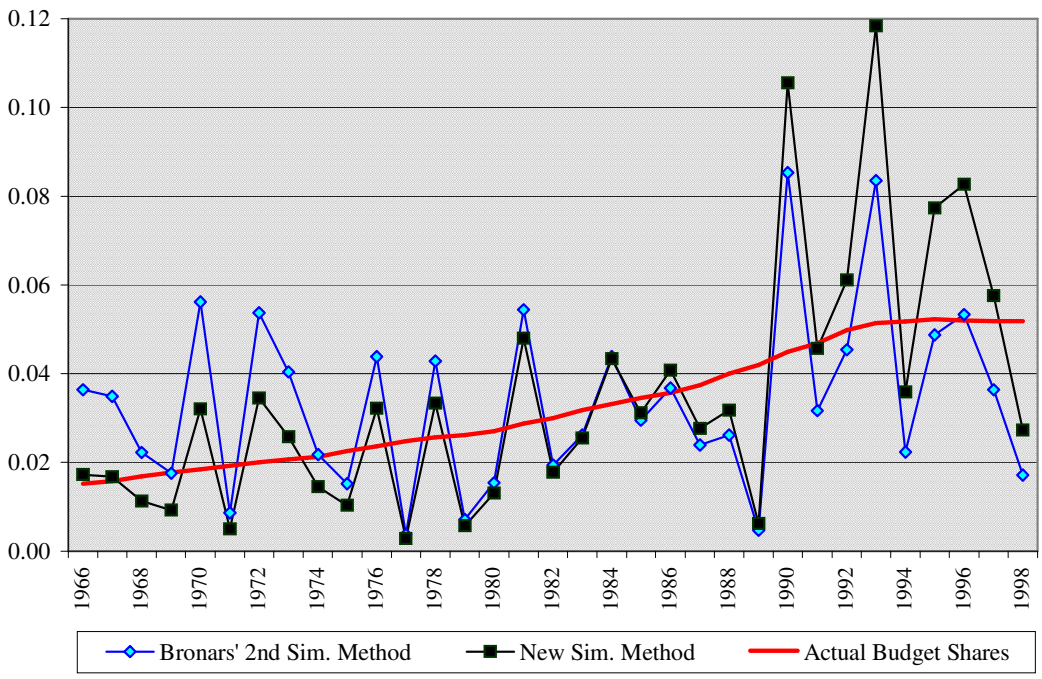

Figure 3.10: Budget shares of Medical Services (S4)
Real and simulated random data, 1966-1998.

Table 3.1: Illustrating the first set of solutions to problem 2

Periods:	T		
	t(1)	t(2)	Annual Average
x_1	200.00	...	...
p_1	100.00	100.00	100.00
x_2	250.00	...	...
p_2	40.00	50.00	45.00
Total Expenditures	30000.00	...	...
Periods:	S		
	s(1)	s(2)	Annual Average
x_1	100.00	...	...
p_1	100.00	100.00	100.00
x_2	500.00	...	...
p_2	50.00	40.00	45.00
Total Expenditures	35000.00	...	...

Table 3.2: Illustrating the second set of solutions to problem 2

Periods:	T			
	t(1)	t(2)	Annual Average	Annual Line Intercepts
x_1	200.00	***100.00***	150.00	300.00
p_1	100.00	84.50	92.25	
x_2	250.00	***425.00***	337.50	675.00
p_2	40.00	42.00	41.00	
Total Expenditures	30000.00	26300.00	27675.00	
Periods:	S			
	s(1)	s(2)	Annual Average	Annual Line Intercepts
x_1	100.00	***200.00***	150.00	270.00
p_1	100.00	120.50	110.25	
x_2	500.00	***40.00***	270.00	607.50
p_2	50.00	48.00	49.00	
Total Expenditures	35000.00	26020.00	29767.50	

Table 3.3: Illustration of a solution to problem 3

Periods:	T		
	t(1)	t(2)	Annual Average
x_1	200.00	0.00	100.00
p_1	100.00	100.00	100.00
x_2	250.00	750.00	500.00
p_2	40.00	40.00	40.00
Total Expenditures	30000.00	30000.00	30000.00
Periods:	S		
	s(1)	s(2)	Annual Average
x_1	100.00	350.00	225.00
p_1	100.00	100.00	100.00
x_2	450.00	0.00	225.00
p_2	50.00	50.00	50.00
Total Expenditures	32500.00	35000.00	33750.00

Table 3.4: GARP consistency of Cobb-Douglas data with and without measurement errors:

The impact of time aggregation

	Proportion of 2000 simulated datasets with no GARP violation				
Magnitude of measurement errors (K):	K=0%	K=1%	K=5%	K=10%	K=20%
1st Cobb-Douglas specification:					
a) 148-observation samples of quarterly data	100.00%	96.80%	20.00%	1.85%	0.00%
b) 37-observation samples of annual (reaggregated) data	100.00%	99.85%	95.65%	87.20%	75.55%
c) 148-observation samples of annual (reaggregated) data, generated from 592 quarterly observations.	100.00%	98.00%	44.30%	13.70%	0.90%
2nd Cobb-Douglas specification:					
a) 148-observation samples of quarterly data	100.00%	93.65%	10.35%	0.50%	0.00%
b) 37-observation samples of annual (reaggregated) data	100.00%	99.70%	93.40%	84.35%	70.45%
c) 148-observation samples of annual (reaggregated) data, generated from 592 quarterly observations.	100.00%	96.75%	33.95%	6.80%	0.30%

Note: The two Cobb-Douglas specifications are: (1) $U(\mathbf{x}) = x_1^{0.60}.x_2^{0.25}.x_3^{0.10}.x_4^{0.04}.x_5^{0.01}$; (2) $U(\mathbf{x})=x_1^{0.40}.x_2^{0.30}.x_3^{0.15}.x_4^{0.10}.x_5^{0.05}$

Table 3.5 - The Power of GARP test using Bronars' (1987) algorithms

		# of Budget Line Intersections			Rejecting H0 (GARP consistency)	
Quarterly Datasets:	Sample	Minimum	Median	Maximum	Method 1 [$S_{it}(1)$]	Method 2 [$S_{it}(2)$]
Q1: Leisure included	(1964:I - 2000:IV)	117	145	147	100.0%	100.0%
Q2: Leisure excluded	(1959:I - 2000:IV)	12	98	165	100.0%	100.0%
Annual Datasets:	Sample	Minimum	Median	Maximum	Method 1 [$S_{it}(1)$]	Method 2 [$S_{it}(2)$]
A1: Leisure included	(1964 - 2000)	29	35	36	99.1%	78.8%
A2: Leisure excluded	(1959 - 2000)	8	25	40	84.1%	68.4%

Note: Each of the last two columns shows percentage of times the null hypothesis was rejected in 2000 simulations of random data.

Table 3.6: Applying Bronars' approach on Cobb-Douglas demand data, with and without measurement errors

		Rejecting H0 (GARP-Consistent) when HA is true			
		Bronars (1987)		Burton (1994)	Cox(1997)
	Error (K%):	$S_{it}(1)$	$S_{it}(2)$	$S_{it}(3)$	$S_{it}(4)$
	0%	95.50%	96.50%	0.00%	0.00%
	1%	96.00%	96.50%	0.50%	0.00%
1st Cobb-Douglas specification	5%	98.00%	96.50%	13.00%	12.50%
coefficients: [0.60;0.25;0.10;0.04;0.01]	10%	96.50%	99.00%	17.50%	22.50%
	20%	96.00%	98.00%	44.00%	42.50%
	0%	98.70%	98.15%	0.00%	0.00%
	1%	98.40%	98.35%	0.15%	0.20%
2nd Cobb-Douglas specification:	5%	98.15%	98.55%	9.80%	10.70%
coefficients: [0.40;0.30;0.15;0.10;0.05]	10%	98.40%	98.60%	24.30%	26.00%
	20%	98.75%	98.40%	43.25%	49.80%

Note: Each of the last 4 columns shows percentage of times the null hypothesis was rejected in 2000 simulations of random data.

Table 3.7 - Model-fitting results[(1)] for consumption subcategories, data on budget shares at multiple frequencies (1964-2000)

Variables[(2)]:	Best-fitting models[(3)]	
	Quarterly data	**Annual data**
Motor vehicles and parts (D1)	ARIMA(4,1,4) F-statistic: 7.4878 Prob(F-statistic): <0.0001	ARIMA([4],1,[4])†,†† F-statistic: 5.8915 Prob(F-statistic): 0.0071
Furniture and household equipment (D2)	IMA(1,1) F-statistic: 15.1835 Prob(F-statistic): 0.0001	ARIMA([4],1,[4]) F-statistic: 8.7959 Prob(F-statistic): 0.0010
Other durables (D3)	IMA(1,1) F-statistic: 21.6435 Prob(F-statistic): <0.0001	RW†
Food (ND1)	ARIMA([4],1,[4])† F-statistic: 5.5971 Prob(F-statistic): 0.0046	RW†
Clothing and Shoes (ND2)	ARI(12,1) F-statistic: 1.7938 Prob(F-statistic): 0.0683	RW†
Gasoline and oil (ND3)	ARIMA(3,1,3) F-statistic: 3.7026 Prob(F-statistic): 0.0019	RW†
Fuel oil and coal (ND4)	ARIMA([4],1,[4])†,†† F-statistic: 8.7064 Prob(F-statistic): 0.0003	IMA(1,1) F-statistic: 6.7570 Prob(F-statistic): 0.0137
Other nondurable goods (ND5)	ARIMA([1,4],1,[1,4])† F-statistic: 7.1983 Prob(F-statistic): <0.0001	ARIMA([4],1,[4]) F-statistic: 10.1776 Prob(F-statistic): 0.0004
Housing (S1)	ARIMA(1,1,1) F-statistic: 17.8857 Prob(F-statistic): <0.0001	ARIMA(3,1,3) F-statistic: 12.4618 Prob(F-statistic): <0.0001
Household Operation (S2)	ARIMA(2,1,2) F-statistic: 9.6170 Prob(F-statistic): <0.0001	ARIMA(1,1,1)†,†† F-statistic: 9.1633 Prob(F-statistic): 0.0048
Transportation (S3)	ARIMA(1,1,1) F-statistic: 16.0796 Prob(F-statistic): <0.0001	ARIMA(1,1) F-statistic: 7.8593 Prob(F-statistic): 0.0084
Medical Care (S4)	ARIMA(12,1)* F-statistic: 8.9852 Prob(F-statistic): <0.0001	ARI(1,1) F-statistic: 48.6093 Prob(F-statistic): <0.0001
Recreation (S5)	ARIMA([4],1,[4]) F-statistic: 9.1555 Prob(F-statistic): <0.0001	ARIMA([4],1,[4]) F-statistic: 11.0168 Prob(F-statistic): 0.0003
Other services (S6)	ARIMA(1,1,1)† F-statistic: 3.8978 Prob(F-statistic): 0.0224	ARI(1,1)†,†† F-statistic: 11.7753 Prob(F-statistic): 0.0003
Leisure (Leis)	RW†	IMA(1,1)† F-statistic: 5.7847 Prob(F-statistic): 0.0218

Notes: (1) Following Rossana and Seater's (1995) notation, we marked with "†" the cases in which the IMA(1,1) specification also generated white-noise residuals, but some other model was preferable according to the Schwartz criterion; similarly, we used "††" to indicate that the Random Walk (RW) specification was acceptable but not preferable; "*" indicates the cases in which no model generated white-noise residuals; (2) Augmented Dickey-Fuller tests under two specifications - with a constant or with constant and presence of a unit root at the 5% significance level, in all cases; (3) Prob(F-statistic) indicates at what significance level one can trend terms - were applied to each series; the tests did not reject the reject the hypothesis that all terms in the regression equation (excluding the constant) are null.

Table 3.8: Annual budget shares, extreme values along 1964-2000

	% of total nominal expenditures			
	with leisure		without leisure	
Goods:	Minimum	Maximum	Minimum	Maximum
D1	0.72	1.34	2.99	4.34
D2	1.12	1.76	3.96	6.65
D3	0.31	0.71	1.21	2.52
ND1	4.91	5.42	14.72	26.00
ND2	1.51	1.86	4.83	8.50
ND3	0.66	1.32	2.01	4.98
ND4	0.07	0.24	0.22	1.30
ND5	1.87	2.70	7.34	8.95
S1	3.42	5.00	14.52	15.90
S2	1.39	2.02	5.95	6.94
S3	0.75	1.42	3.35	4.32
S4	1.44	5.22	5.28	16.20
S5	0.51	1.32	2.05	3.94
S6	2.32	5.41	9.65	16.11
LEIS	66.43	77.73	-	-

Table 3.9 - The Power of the GARP test against each alternative hypothesis

		Rejecting H0: Data is GARP-Consistent				
		Bronars (1987)		Burton (1994)	Cox(1997)	New Method
Datasets:	Samples	$S_{it}(1)$	$S_{it}(2)$	$S_{it}(3)$	$S_{it}(4)$	$S_{it}(5)$
Q1: Leisure included	(1964:I - 2000:IV)	100.0%	100.0%	98.9%	96.9%	100.0%
Q2: Leisure excluded	(1959:I - 2000:IV)	100.0%	100.0%	98.9%	96.5%	100.0%
A1: Leisure included	(1964 - 2000)	99.1%	78.8%	6.7%	19.7%	90.5%
A2: Leisure excluded	(1959 - 2000)	84.1%	68.4%	10.1%	28.0%	80.6%

Note: Each of the last 5 columns shows percentage of times the null hypothesis was rejected in 2000 simulations of random data.

Chapter 4

GARP Consistency and Weak Separability of Macroeconomic Aggregates: Accounting for the Consumption of Household Services

4.1 Introduction

In this chapter we apply Varian's (1982, 1983) tools of revealed preference analysis to a dataset recently developed to incorporate the value of nonmarket household services into NIPA'S accounting of U.S. personal consumption expenditures. By doing so, we combine efforts from two research areas that have remained parallel up to this point, with mutual benefits. On the one hand, we contribute to the literature on revealed preference analysis of macroeconomic aggregates in two ways. First, by considering this additional consumption category, we explicitly investigate the so far maintained assumption that all other goods are weakly separable from those nonmarket services in a representative utility framework; we also improve upon previous descriptions of time allocations by distinguishing between time as a source of utility (leisure) or as an input in the household production. On the other hand, we take a necessary step preceding the use of those new statistics in empirical macroeconomic models: as we submit the expanded dataset to the GARP test, we are able to investigate whether there is indeed a well behaved utility function rationalizing those consumption choices.

Most of previous studies (e.g. Varian (1982), Bronars (1987) and FHS) have focused on the representative consumer's expenditures on market goods and services only – assuming *a priori* that those were weakly separable from all other sources of utility. Others incorporated leisure choices based on estimates of the residual amount of time not allocated to work or to sleeping and eating (Swofford and Whitney, 1987, 1988, 1994; Drake, 1997; chapters 1 and 3 of this dissertation). Nevertheless, recent estimates of the value of activities carried out inside the household also permit incorporating the consumption of household services in the revealed preference analysis of macro data. We shall argue that doing so is critical when one investigates the rationality of consumers' choices over the second half of the 20th century.

This chapter is organized in four additional sections. In section 4.2 we discuss how the incorporation of these services is relevant from both theoretical and empirical standpoints. Next we summarize sources and methods of the new data in section 4.3. Then we implement the GARP test, check its power over different subsets of data and study the weak separability of particular aggregates in section 4.4. Section 4.5 has our conclusions.

Our test results indicate that previous conclusions supporting overall GARP-consistency and mutual weak separability of standard annual aggregates are unaffected by the inclusion of household services, despite the changing share of resources allocated to those activities over our 34-year sample. Furthermore, the changing number of hours spent on average in household production – due to the increasing participation of women in the civilian labor force over those years – can be characterized as a rational decision made by the representative agent in a standard utility maximization model. Further efforts on the development of statistics at higher frequencies seem appropriate, though, due to the observed low power of the GARP test within the available annual series.

4.2 Incorporating the consumption of household services

There are at least three reasons to believe that the explicit consideration of this additional consumption category is indeed relevant in testing the existence of a well-behaved representative utility function rationalizing subsets of macro data.

The first reason is purely theoretical. Running the GARP test on a dataset that does not contain a series for a particular good implies maintaining the assumption that all goods considered are weakly separable from the omitted one (Swofford and Whitney, 1988)[75]. Therefore, as estimates for previously unobservable choices are made, the researcher can verify whether GARP violations and/or rejections of specific separability structures within former datasets can be attributed to the exclusion of that good.

The second reason is the very fact that standard statistics on national accounting and personal consumption expenditures do not include estimates of the unpaid production of goods and services inside the household. The consumption of those goods clearly involves nonnegligible amounts of resources and its evolution in recent decades is closely related to the changing participation of women in the civilian labor force (Landefeld and McCulla, 2000). By including this additional consumption category in the revealed preference analysis of aggregates, we investigate a more complete set of choices made by consumers and also check whether recent changes in the average time spent on market vs. nonmarket activities can be characterized as rational decisions of the utility-maximizing representative consumer[76].

[75]Varian has shown – in a somewhat neglected paper (Varian, 1988) – that an incomplete dataset may actually not pass GARP because one of its components is not weakly separable from some excluded good(s). According to Social Science Citation Index (ISI's Web of Science), by July/2004 Varian's (1982, 1983) original papers had been cited at least ten time more often (163 and 107 times, respectively) than his work suggesting a more careful interpretation of the GARP test results (10 hits) due to the fact that researchers observe incomplete sets of prices and quantities associated with consumers' choices.

[76]Despite the well-know caveats associated with the use of representative agent models in macroeconomics – particularly the precariousness of welfare analysis based on those models (Kirman, 1992) –, we share FHS's perception that those are still useful if they can explain aggregate data and/or permit making predictions.

A third justification for the expansion of the relevant dataset involves the manner in which estimates of the enjoyment of leisure have been incorporated into this sort of analysis so far. We believe it is quite appropriate to reevaluate previous calculations of leisure quantities, which relied essentially on the residual amount of time not allocated to paid work or to biological imperatives such as sleeping and eating. Swofford and Whitney (1987, 1988, 1994), Drake (1997) and Drake et al. (2003) assumed a fixed 10-hour daily period for sleeping/eating and calculated leisure time by subtracting the average number of worked hours from the remaining 14 hours[77]. It is not hard to see that such a description of consumers' behaviors neglects a potentially important fourth use of time, spent on unpaid activities such as house cleaning, personal health care and meal preparation.

Fortunately, the new numbers on household services are calculated from estimates of the average time allocated for those activities, and the very series on time use by household members are also available (see next section). Therefore, the expanded dataset permits a more detailed description of the allocation of time in a given period and particularly, a more precise picture of how leisure choices have evolved over our sample. Simply put, as the average number of hours spent on household work is reduced, the amount of (residual) leisure time will tend to grow faster than previously assumed.

4.3 Valuing unpaid household services and recalculating leisure time

In this section we summarize Landefeld and McCulla's (2000) efforts to expand Eisner's (1989) annual series on the value of household services for the period 1946-1997, consistently with existing NIPA tables of marked-based transactions[78]. We rely

[77]In a different sort of empirical study, Mankiw et al. (1985) also assumed a daily fixed amount of time (only 8 hours, though) allocated neither to work nor leisure.

[78]For details on all other data used in this chapter – discussed in chapter 1 –, see our Appendix. Recall that our analysis is restricted to samples starting in 1964, due to the limited span of (consistent) leisure data.

mostly on a survey of previous attempts made by the latter (Appendix A in Eisner (1989)) and on detailed tables used in the construction of more recent estimates, provided by the former. The use of these new series has advantages and shortcomings, which are discussed in the next two subsections, respectively.

4.3.1 The fourth use of time and its possible impacts on previous findings

Landefeld and McCulla (2000) pointed out that the debate about valuation and inclusion of household services in national income accounting is not new; it actually goes back to the beginning of national accounting, having received special attention in the 1970s with a surge of interest in welfare accounting[79]. They modified the NIPA tables to include household services as a component of personal consumption expenditures (and therefore, of GDP) and to treat the acquisition of consumer durables as investment, or inputting the value of services that stocks are expected to provide over time[80]. Their work improved upon previous attempts to adjust the NIPA tables for having first disaggregated household production in a I-O framework, *"…thus allowing consideration of the relation between households and other industries and the economy as a whole"* (Landefeld and McCulla, 2000).

The first aspect above is of main interest here. The particular approach used to estimate the value of household services had been adopted before by Kendrick (1976). Labeled "market alternative=housekeeper cost" (MAHC) by Hawrylyshyn (1976), it involves multiplying the time spent by household members on those activities by the

[79]Previous studies accounting for broader ranges of economic activities include Ruggles and Ruggles (1970, 1982), Nordhaus and Tobin (1972, 1973), Kendrick (1976), Zolotas (1981), Jorgenson and Fraumeni (1987).

[80]Notice that the second aspect motivating their paper is actually a standard practice in the empirical literature of GARP-consistency of macroeconomic aggregates, which has always adopted estimates for the user costs of stocks of durables as the relevant expenditure on those consumption categories (rather than the current purchase of new equipment published in NIPA).

before-tax wage rate of general-purpose domestic workers performing the same duties[81]. The resulting figure is an estimate of how much the market-substitute for those services would cost, on an hourly basis.

Hawrylyshyn (1976) identified two alternatives for the valuing of hours spent on household work. The first one, merely a variation of the method just described, involves multiplying the time spent on each household activity by the market wage rate for specific functions, being labeled "market alternative=individual function cost" (MAIFC) for that reason. Estimating the opportunity cost of household members' time constitutes the second alternative approach, which was actually implemented by Nordhaus and Tobin (1973) and Zolotas (1981), among others. Based on Becker's (1965) description of the optimal allocation of time by households, this last method is referred to as "wage=opportunity cost of time"[82]. Despite concerns expressed by Hawrylyshyn that the use of specific methods are more likely to under/overestimate the value of those services, Chadeau's (1985) surveyed the applications of the three different approaches – in multiple nations – and concluded that the final numbers are, overall, remarkably consistent.

The numbers of hours allocated to household services were calculated as weighted averages of time spent by male/female, employed/unemployed household members on those activities – the original data on time use coming mostly from surveys performed by the Inter-University Consortium for Political and Social Research (1979, 1983) at the University of Michigan[83]. Eligible respondents were between 18 and 65 years of age, members of households with at least one employed adult in a nonfarming occupation[84].

---

[81]At least a part of the series on wage rates of domestic workers constructed by Eisner (1989) for the period 1946-1981 involved the use of unpublished data provided by Robert Parker, from the Bureau of Economic Analysis. Nevertheless, the whole series is built to reflect the hourly total compensation of employees in private households. See Eisner (1989), pp. 66-68.

[82]This last method is actually standard in the estimation of the value of leisure hours; see our Appendix.

[83]Landefeld and McCulla(2000) relied on comparable numbers produced by Robinson and Godbey(1997).

[84]For more details on surveys criteria and on data interpolation procedures, see Eisner (1989) pp.57-66.

Figure 4.1 plots the series on per capita average number of hours allocated to nonmarket activities, with and without the explicit consideration of household services. Leisure is always calculated as the residual amount of time not allocated to any other activity. The highest and flattest line in the figure represents the amount of time presumably allocated to leisure, in accordance with all previous works (aforementioned) that do not make the distinction between leisure and household services[85]. As one introduces the distinction between the two nonmarket activities, though, that initial estimate of leisure time is divided into two components with markedly distinct evolutions over time. By the beginning of our sample, the representative consumer allocates – by coincidence – approximately the same amount of time (about 4 hours/day, on average) to unpaid work and leisure; three decades later, the average number of working hours inside the household is reduced to 3.6/day, as 5.5 hours are allocated for daily leisure time[86].

To stress one last time a central issue discussed here, notice that previous estimates of the total amount of leisure time is not simply downward-shifted with the explicit consideration of household work; its trend was actually underestimated. Rather than increasing 6.9% (from 8.5 to 9.1 hours/day) over those 34 years, the amount of leisure time has actually grown 28.8 % (from 4.3 to 5.5 hours a day, approximately), as we consider the (15.3%) reduction in the number of hours spent on household services. Therefore, neglecting this fourth use of time has had a significant impact on one's perception of the evolution of leisure choices and may have, in theory, affected previous conclusions on their weak separability from other consumption aggregates.

Intuitively, one is more likely to find evidence that two goods are separable if the quantities demanded of one of them does not seem to be affected by changes in their

[85]The list of activities includes cooking, cleaning, laundry, management (paperwork), animal/plant care, repairs, yard (outdoor work), child care, adult healthcare and shopping, among others. Although some of the activities can clearly be considered enjoyable ones or hobbies, it is hard to believe that performing each and all of those brings direct satisfaction to any given person.

[86]Male readers not convinced that the average amount of time spent on leisure has actually grown over the last few decades should try convincing their wives to quit their jobs and go back home to wash/iron the

relative price. As previous findings were based on such an imperfect description of time allocations over those years, the results may have been biased toward the nonrejection of weak separability of leisure choices.

In sum, our analysis does not simply add to previously studied datasets a different consumption subcategory, from which standard macroeconomic aggregates have been implicitly assumed weakly separable in all previous studies; we are also improving upon all prior descriptions of leisure choices.

4.3.2 Limitations of available data on nonmarket services

Accounting for the use of time is undoubtedly a difficult task for many reasons. Researchers must rely on surveys that actually access information on people's perception of their time allocations, which may quite often be imprecise[87]. Nevertheless, of most concern at this point is the fact that long samples of data on the value of time spent with household production are only available on an annual basis. As discussed throughout the previous chapter, the unavailability of data at higher frequencies may be an significant limitation to the analysis of particular consumption choices because the GARP test is likely to have low power in those cases[88]. Therefore, it is wise to investigate also the power of the test in the expanded dataset.

As also discussed in chapter 3, prior knowledge of the behavior of budget shares over the studied sample permits a more careful interpretation of the results from Bronars' (1987) approach – particularly when the outcomes seem to be sensitive to the

---

clothes, clean the house, water the plants and cook, all before 5:30 pm when they will finally have the privilege of… taking care of their husbands needs.

[87] Robinson and Godbey (1997) – Landefeld and McCulla's source of statistics on time use – has relied on "time diaries" in an attempt to keep track of people's activities more precisely than previous after-the-fact interviews. A second concern is that researchers must interpolate data for the periods between surveys, as there is no continuous gathering of information on those phenomena.

[88] We refer to the fact that test fails to reject the null hypothesis (GARP-consistency) in randomly simulated annual data much more often than with quarterly figures generated from the same "irrationality"

adoption of alternative irrationality models. Recall that among 5 alternative data-generating algorithms, we found that two of them – Burton's(1994) and Cox's (1997) – underestimated the power of the test relative to the others, with divergent results being attributed to the fact that the budget shares for multiple consumption subcategories varied little over the sample[89].

Figure 4.2 shows that the value of the household production remained a stable proportion (around 4%) of the total expenditures – which included also the purchases of market nondurable goods and services, the user cost of durables and estimates for the value of leisure time. Notice also that the evolution of leisure shares basically shifted downwards with the introduction of household services in the dataset, remaining the largest expenditure category for the entire sample – exceeding half of total expenditures[90].

Those observations suggest that the use of Burton's or Cox's simulation algorithms to evaluate the power of the test may once more be inadequate; if the actual data passes GARP, the test will tend to detect no inconsistencies in the "random" series generated with those methods; the test will (most likely) not discern between actual figures and supposedly irrational ones – simulating the consumption of household services – simply because they are very similar, by construction. In other words, both methods are expected to fail in generating irrational choices under the circumstances, precisely as it happened in chapter 3.

As one compares figure 4.1 and 4.2, it is worth noticing – for reasons that will be evident soon – that even though the amount of time spent on household activities has clearly decreased over those years (figure 1), its total value remained almost a constant

---

models. If the GARP test rejects half of the simulated datasets at a certain frequency, we say that the power of the test against the specific model of random behavior generating those figures is 50%.

[89]Both algorithms generate random budget shares that are limited inside the ranges of observed figures; since actual budget share allocations for many consumption subcategories varied little, simulated and actual figures tended to behave very similarly.

[90]Except from the decreasing consumption of fuel oil and coal (ND4) over those years, real expenditures on all other goods and services have grown faster than the corresponding figures for household services and leisure; that was clearly the most important factor driving down the share of leisure in total expenditures, represented by its overall negative trend in figure 4.2.

share of expenditures over the entire sample (figure 2). In theory, changes in relative prices and/or productivity gains can explain those circumstances. Consumers may be actually producing more household services using fewer hours per day, as new/affordable technologies have possibly become available. Alternatively or concomitantly, changes in the opportunity cost of time (wage rates) could have put upward pressure on consumers' willingness to pay for the market substitutes. Nevertheless, since there is no series on final prices for the activities performed by households – as observable in market transactions – the identification of the true sources of changes in nominal expenditures becomes impractical, if not impossible.

To further understand this last assertion, consider the textbook story supporting the incorporation of household production into national accounts, according to which a guy marries his cook /housekeeper and, as a result, GDP is reduced – as those services are still provided but not accounted for in standard statistics. 10 years passed, the cook now spends slightly less time performing those services. In the job market, the wage rate for employees in that industry has doubled (fictitious figures). As we try to infer how much of those services the household consumes now, we simply cannot know precisely: it could be the case that the cook is much more productive than when she quit her job and that, say, 50% more services are produced within about the same number of hours; but it might also be true that the marriage is not going swell, and the guy is actually risking his life by eating at home. The main point is that we have no information about productivity in the household sector, which can be different from that of a professional cook/housekeeper.

The discussion above reveals what can be considered a second caveat, besides the data frequency, of using the available estimates on the value of nonmarket activities in this sort of analysis. We refer to the fact that real expenditures on household services were calculated by deflating nominal values not by an index of final prices (which simply do not exist) but by that of an input: the wage rate of domestic workers performing those duties in the job market – following Eisner (1989). It has been argued that such a procedure "…*would result in low or zero productivity in the household sector*

and bias real growth in household relative to marked production." (Landefeld and McCulla, 2000, pp.300, fn.#13)[91]. Since Varian's framework adopted throughout this dissertation requires some sort of decomposition of changes in expenditures into prices and quantities movements, and for lack of better alternative, we followed Eisner and converted the current-value estimates into 1996 constant dollars[92].

4.4 Test results

Recall that if an overall dataset passes GARP and, say, nondurables pass both conditions for weak separability – as proposed by Varian (1983) and implemented by Fleissig and Whitney (2003)[93] – there is a well-behaved representative utility function rationalizing that subset of the data. On the other hand, if either the overall dataset or a specific combination of consumption categories does not pass the necessary condition (GARP-consistency), one has evidence that a maximization model set only on those goods is in fact misspecified[94].

As for the GARP-consistency of the expanded dataset (including the consumption of household services and adjusting the real amount of leisure time actually enjoyed by the representative consumer), we found not a single GARP violation; therefore, the overall dataset can indeed be rationalized by a representative agent model.

---

[91]This was actually their justification for not publishing real estimates of their adjusted national accounts; all tables were exclusively presented and discussed in current values.

[92]We believe that Eisner's procedure actually entails the implicit assumptions that the husband would be willing to pay his wife for those services as much as she could get in a formal employment contract, and that the lady would be equally productive in household and market sectors over time.

[93]Besides Varian's seminal paper itself, readers are referred to Fleissig and Whitney (2003) for a good discussion on the algorithms checking whether Afriat inequalities hold for specific data subsets.

[94]Fleissig, Gallant and Seater's (2000) have shown that the aggregation method used to combine a subset of goods does matter for the revealed preference analysis of aggregates and for further investigations of the model's empirical performance; rather than using the simple sum of real expenditures on nondurables and services, as the common practice, researchers should aggregate those consumption subcategories using superlative indices before actually testing a model's implications. This recommendation in fact has become a "rule" with BEA's adoption of chained-indices in the calculation of real figures; see chapter 2 for a detailed discussion on new data methodologies.

To verify the robustness of the conclusion above, the power of the test is checked in datasets with and without household services, A1 and A2, respectively (table 4.1); as in chapter 3, we started checking how often budget lines intersect, as the test can be expected to have very little power in datasets that contain few budget intersections. The first three columns show minimum, median and maximum number of times the budget hyperplane for any give year intersects with those for other periods. The next five columns contain rejection rates – in percentage terms – that show how often series of random data did not pass GARP, out of 2000 total simulations with each algorithm[95]. Recall that each of those five methods involves generating series on budget shares that, although random, preserved specific characteristics of the actual figures, as very briefly summarized below[96]:

S_{it}(1): Proposed by Bronars (1987), simulated budget shares are basically drawn from a uniform random distribution; simulated real expenditures (quantities) are calculated with the product of those random budget shares and total nominal expenditures, subsequently divided by the actual prices.

S_{it}(2): Also Bronars', the series on simulated budget shares are constructed in such a way that figures fluctuate randomly around the historical average of actual shares.

S_{it}(3): Burton's (1994): series on simulated budget shares fluctuate randomly inside the ranges of values actually observed over the sample.

S_{it}(4): Cox's (1997) method randomly reassigns actual budget shares for all periods, maintaining the series on prices in the original order.

[95]Also as in the preceding chapter, the simulations were run over datasets that assumed two different expectation schemes in the calculation of user costs of durable goods; as results were (again) not sensitive to such assumptions, the only figures reported here are the ones for the case of perfect foresight, for simplicity.

[96]For more details on any of the algorithms, refer back to section 3.3 of this dissertation.

$S_{it}(5)$: Our own method is a variation of $S_{it}(2)$: rather than fluctuating randomly around the overall sample average, simulated budget shares oscillate from local (moving) averages of actual figures.

Inspection of the numbers in table 1 reveal, first, that the introduction of household services did not significantly affect the frequency of budget intersections: budget hyperplanes for any given year intersected with at least 22 others within the dataset excluding household services, and at least 23 times as the new category was incorporated. In both cases, though, at least half of the budget hyperplanes intersected with all other ones (for all years).

As we ran the GARP test on random data generated with each algorithm, the same pattern observed in chapter 3 emerged from both datasets A1 and A2: (i) Bronars's first algorithm indicated a very high power of the test – about 95% – but against what has been considered an extreme model of irrationality (see chapter 3); (ii) as random budget shares were built to fluctuate around their corresponding historical averages [$S_{it}(2)$], the test failed to reject the null hypothesis quite often (almost 40% of the time in the expanded dataset); (iii) Burton's (1994) and Cox's (1997) simulation algorithms still led to an underestimation of the power of the test relative to all other methods[97] – although with almost twice as high rejection rates in the expanded dataset; (iv) the use of the new algorithm proposed in chapter 3 indicated that the power of the test was not as high as observed with Bronars' first method, but not as low as one would find using his second algorithm. The null hypothesis was rejected about 80% of the time.

Overall, we interpret the results above as evidence that the power of the test was not increased with the explicit consideration of choices regarding the consumption of household services and that, analogous to our findings in chapter 3, further investigations on currently unavailable high-frequency statistics may indicate that there

[97]Such finding is not surprising, since expenditures on the new consumption category constituted almost a fixed proportion of total expenditures over those years, following the pattern of other subcategories (chap.3).

are no well-behaved utility functions rationalizing the most commonly adopted subsets of data at the quarterly or monthly frequency.

Since the hypothesis of GARP consistency was not rejected in our annual dataset, we proceeded to check the weak separability of preference structures commonly assumed in representative agent models. Table 4.2 shows that all major consumption aggregates passed necessary and sufficient conditions for mutual weakly separability at the annual frequency, regardless of the explicit consideration of the consumption of household services. The first half of the table simply confirms – over a subsample of the annual data previously studied (1964-2000) – the findings on weak separability of preference structures discussed in chapter 1: researchers can set and investigate models on different combinations of annual consumption aggregates, including or not the series on leisure choices; our tests could not reject the existence of well-behaved representative utility functions rationalizing many subsets of consumption categories.

Surprisingly, to some extent, the same conclusion was reached within the expanded dataset, including the annual consumption of household services (lower half of table 4.2). As before, the most commonly assumed preference structures – set on nondurables and services or on nondurables only – were not rejected by our tests; however, some other interesting structures also passed both necessary and sufficient conditions for weak separability. Take for example SEP6(a), which involves testing whether household services along with the other market services can be considered a weakly separable aggregate from all other goods. As this proposed structure passed the GARP test (the necessary condition for weak separability), we did not reject the existence of a well behaved representative utility function rationalizing the set of 6 subcategories of services and the seventh nonmarket counterpart. Subsequently, the seven subcategories passed the Afriat conditions for weak separability from all other disaggregated categories of consumption goods, including leisure. As a result, empirical researchers interested in studying the demand for all services can test the implications of a maximization model set and solved for those annual consumption categories only –

conditional only on the total expenditure allocated to the overall aggregate and not on a maintained assumption about the separability of those goods in the utility function.

Another result with particularly appealing interpretation is the weak separability of both nonmarket activities (leisure and household services) from each other and from all other annual consumption aggregates (SEP1, SEP5). It implies, intuitively, that household members decide first how much time will be spent on leisure vs. household services, before breaking up unpaid work time among tasks such as house cleaning or cooking.

4.5 Final remarks and conclusions

In this paper we improved upon previous studies on GARP consistency and weak separability of macroeconomic aggregates in two ways. As recent estimates of household services were incorporated for the first time in this sort of analysis, we directly accounted for a nonnegligible share of consumers' budgets. The omission of that aggregate in all previous work makes them investigations of the maximizing behavior conditional on a strong maintained assumption: that all sets of goods, services and leisure considered before were weakly separable from the services produced inside the household. As that consumption category was incorporated, we were able to test this hypothesis, finding that it was not rejected at the annual frequency.

Additionally, the expanded dataset studied here also constituted a better description of consumers' choices because it permitted a more accurate estimation of the average amount of leisure enjoyed by consumers over time. As women's participation in the labor force increased in recent decades, there was a significant reduction in the average number of hours spent on household services. Accounting for this reallocation of time from nonmarket (unpaid) work to leisure, we found that the value of household services remained a stable proportion of the consumers' total expenditures – in per capita

terms, including also purchases of new nondurables and services, the user costs of durable goods and the opportunity cost of residual time, presumably used for leisure.

Our test results indicate that commonly adopted subsets of consumption categories can be rationalized by well-behaved representative utility functions at the annual frequency. They also support the view that major reallocations of the average use of time by households can be indeed characterized as rational decisions made by a representative consumer. Nevertheless, these conclusions should not be extended automatically to datasets of higher frequency; further development of quarterly statistics on the value of household production, as well as estimates on the productivity of the household sector, may still have significant impacts on the GARP-consistency and weak separability of subsets of data.

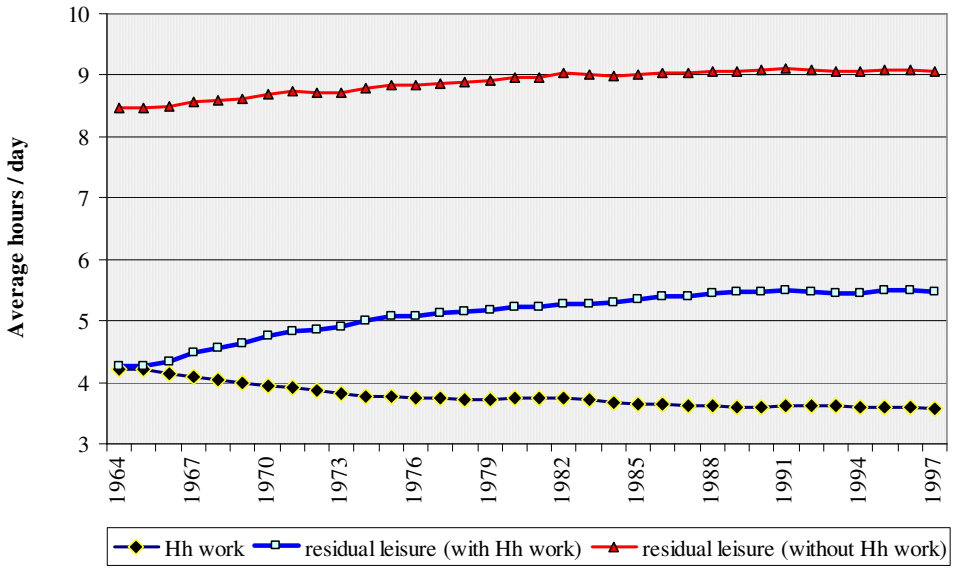

Figure 4.1: Time allocation, 1964-1997
Breaking up previous estimates of per capita average numbers of hours allocated to nonmarket activities into leisure and household work.

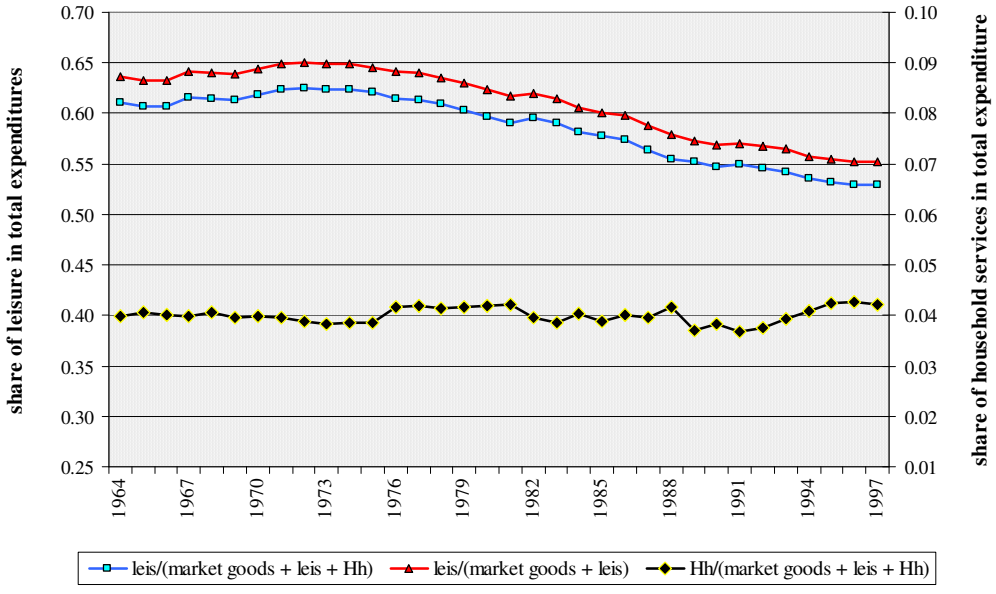

Figure 4.2: Evolution of budget shares
Datasets with and without household services (Hh)

Table 4.1 - The power of the GARP test against distinct alternative hypotheses

	# of budget intersections	rejecting H0: data are GARP-consistent				
		Bronars (1987)		Burton(1994)	Cox(1997)	Maia(chap.3)
Datasets:	[min , median , max]	$S_{it}(1)$	$S_{it}(2)$	$S_{it}(3)$	$S_{it}(4)$	$S_{it}(5)$
A1: Household services included	[23 , 33 , 33]	95.5%	61.6%	12.5%	21.4%	77.2%
A2: Household services excluded	[22 , 33 , 33]	95.9%	68.2%	5.3%	13.7%	80.5%

Note: Each of the last 5 columns shows percentage of times the null hypothesis was rejected in 2000 simulations of random data.

Table 4.2: Weakly separable macroeconomic aggregates, 1964-1997

Preference structures passing both necessary and sufficient conditions for weak separability, as in Varian's (1983) Theorem

Preference structures on subsets of annual data **without HH**:

SEP1	$U [u_1(D) , u_2(ND) , u_3(S) , u_4(Leis)]$
(a)	$U [u_1(D) , ND , S , Leis]$
(b)	$U [u_1(ND) , S , D , Leis]$
(c)	$U [u_1(S) , ND , D , Leis]$
SEP2	$U [u_1(ND,S) , u_2(D) , Leis]$
(a)	$U [u_1(ND,S) , D , Leis]$
SEP3	$U [u_1(D,ND,S) , Leis]$
SEP4	$U [u_1(ND,S,Leis) , D]$

Preference structures on subsets of annual data **including Hh**:

SEP1	$U [u_1(D) , u_2(ND) , u_3(S) , u_4(Leis), u_5(Hh)]$
(a)	$U [u_1(D) , ND , S , Leis , Hh]$
(b)	$U [u_1(ND) , S , D , Leis , Hh]$
(c)	$U [u_1(S) , ND , D , Leis , Hh]$
SEP2	$U [u_1(ND,S) , u_2(D) , Leis , Hh]$
(a)	$U [u_1(ND,S) , D , Leis , Hh]$
SEP3	$U [u_1(D,ND,S) , Leis , Hh]$
SEP4	$U [u_1(ND,S,Leis) , D , Hh]$
SEP5	$U [u_1(Leis,Hh) , D , ND , S]$
SEP6	$U [u_1(S,Hh) , u_2(ND) , D , Leis]$
(a)	$U [u_1(S,Hh) , D , ND , Leis]$

Note: ND, S, D, Leis, Hh stand for disaggregated sets of nondurables (ND1,…, ND5), services (S1,…,S6), durables (D_1,..,D_3), leisure and household services, respectively;

References

Afriat, S. (1967). "The construction of a utility function from expenditure data", *International Economic Review* **8**, 67-77.

Afriat, S. (1972). "Efficiency estimates of production functions", *International Economic Review* **8**, 568-598.

Aizcorbe, A.M.(1991). "A lower bound for the power of nonparametric tests", *Journal of Business and Economic Statistics* **9**(4), 463-467.

Alston, J.A. and J.M. Chalfant (1988). "Accounting for changes in tastes" Journal of Political Economy **26**(2), 391-410.

Barnett, W.A. (1979). "The joint allocation of leisure and goods expenditures", *Econometrica* **47**(3), 539-563.

Barnett, W.A. and S. Choi (1989). "A Monte Carlo study of tests of blockwise weak separability" *Journal of Business and Economic Statistics* **7**, 363-377.

Barnett, W.A., Fisher, D., and A. Serletis (1992). "Consumer theory and the demand for money", *Journal of Economic Literature* **4**, 2086-119.

Becker, G.S. (1962). "Irrational behavior and economic theory", *Journal of Political Economy* **70**,1-13.

_____ (1965). "A theory of the allocation of time", *Economic Journal* **75**(299), 493-517.

Bell, W.R. and S.C. Hillmer (1990). "The time series approach to estimation for repeated surveys", *Survey Methodology* **16**(2),195-215.

References (continued)

Blackorby, C., D. Primont and R. Russell (1998). "Separability: A Survey", in: *The Handbook of Utility Theory Vol. 1*, Barbera, S., P. Hammond and C. Seidl (eds.). Kluwer: Dordrecht.

Braithwait, S.D. (1980). "The substitution bias of the Laspeyres price index: an analysis using estimated cost-of-living indices." *American Economic Review* **70** (1), 64-77.

Bronars, S.G. (1987). "The power of nonparametric tests of preference maximization", *Econometrica* **55**, 693-698.

Browning, M. (1989). "A nonparametric test of the life-cycle rational expectations hypothesis", *International Economic Review* **30**, 979-992.

Burton, M. (1994). "The power of non-parametric demand analysis when applied to British meat and fish consumption data", *European Review of Agricultural Economics* **21**, 59-71

Campbell, J. and N.G. Mankiw (1989). "Consumption, income and interest rates: reinterpreting the time series evidences" in: *NBER Macroeconomics Annual 1989*, Blanchard O.J. and S. Fischer, eds., Cambridge: MIT Press.

Chadeau, A. (1985). "Measuring household activities: some international comparisons", *Review of Income and Wealth* **31**(3), 237-253

Cox, J.C. (1997). "On testing the utility hypothesis", *Economic Journal* **107**, 1054-1078.

Davidson J.E., D.F. Hendry, F. Srba and S.Yeo (1978). "Econometric modeling of the aggregate time-series relationship between consumer's expenditures and income in the United Kingdom", *Economic Journal* **88**, 661-692.

References (continued)

Diewert, W.E. (1974). "Intertemporal consumer theory and the demand for durables," *Econometrica* **42**, 497-516.

_____(1976). "Exact and superlative index numbers", *Journal of Econometrics* **4**, 115-145.

Drake, L. (1997). "Nonparametric demand analysis of U.K. personal sector decisions on consumption, leisure and monetary assets: A reappraisal" *Review of Economics and Statistics* **79**(4), 679-683

Drake, L. , Fleissig, A.R. and J.L. Swofford (2003). "A semi-nonparametric approach to the demand for UK monetary assets" *Economica* **70**, 99-120.

Enders, W. (1995). *Applied Econometric Time Series*, New York: John Wiley and Sons.

Eisner, R. (1989). *The Total Incomes System of Accounts*, Chicago: The University of Chicago Press, Ltd.

Famulari, M. (1995). "A household-based nonparametric test of demand theory", *The Review of Economics and Statistics* **77**(2), 372-382.

Fisher, D. and A. Fleissig (1997). "Monetary Aggregation and the Demand for Assets", *Journal of Money, Credit and Banking* **29** (4), 458-475.

Fleissig, A.R. (1993). *Durability and nonseparability in consumption*, *Unpublished doctoral dissertation*, North Carolina State University.

Fleissig, A.R., Hall, A. and J.J. Seater. (2000). "GARP, Separability, and the Representative Agent." *Macroeconomic Dynamics* **4**(3), 324-342.

References (continued)

Fleissig, A.R., Gallant, R. and J. J. Seater (2000). "Separability, Aggregation, and Euler Equation Estimation", *Macroeconomic Dynamics* **4**(4), 547-572.

Fleissig, A.R. and G. A. Whitney (2003). "A new pc-based test for Varian's weak separability conditions", Journal of Business and Economic Statistics **21**(1), 133-144.

Gross, J.(1995). "Testing data for consistency with revealed preference" *The Review of Economics and Statistics* **77**, 701-710.

Hahm, J.H. (1998). "Consumption adjustment to real interest rates: Intertemporal substitution revisited" *Journal of Economic Dynamics and Control* **22**(2), 293-320.

Hamilton, J.D.(1994). *Time Series Analysis*. Princeton, NJ: Princeton University Press.

Hansen, L.P. and K.J. Singleton (1982). "Generalized instrumental variables estimation of non-linear rational expectations models", *Econometrica* **50**, 1269-86.

Hawrylyshyn, O. (1976). "The value of household services: a survey of empirical estimates." *Review of Income and Wealth* **22**(2), 101-131.

Heaton, J. (1995). "An empirical investigation of asset pricing with temporally dependent preference specifications", *Econometrica* **63**(3), 681-717.

Holman, J. A.(1998). "GMM Estimation of a Money-in-Utility-Function Model: The implications of Functional Forms", *Journal of Money, Credit and Banking* **30**(4), 679-699.

Inter-University Consortium for Political and Social Research (1979). *Time use in economic and social accounts, 1975-1976* . Ann Arbor, MI: Survey Research Center, Institute for Social Research.

References (continued)

Inter-University Consortium for Political and Social Research (1983). *Time use longitudinal panel study, 1975, 1981.* Ann Arbor, MI: Survey Research Center, Institute for Social Research.

Jorgenson, D.W. and B.M. Fraumeni (1987). "The accumulation of human and non-human capital, 1948-1984" Unpublished Manuscript, Harvard University.

Katz, A.J. and S.W. Herman (1997). "Improved estimates of fixed reproducible tangible wealth, 1929-95", *Survey of Current Business* **77**(5), Washington, D.C.: U.S. Department of Commerce.

Kendrick, J.W. (1976). *The formation and stocks of total capital*. New York: Columbia University Press.

Kirman, A.P. (1992). "Whom or what does the representative individual represent?" *Journal of Economic Perspectives* **6**, 117-36.

Kocherlakota, N.R. (1996). "The equity premium: it's still a puzzle", *Journal of Economic Literature* **34**, 42-71.

Landefeld, J.S. and S. McCulla (2000). "Accounting for nonmarket household production within a National Accounts Framework", *Review of Income and Wealth* **46** (3), 289-307.

Landefeld, J. S. and R. Parker(1997). "BEA's Chain Indexes, Time Series, and Measures of Long-Term Economic Growth" *Survey of Current Business* **77**(5), 58-68.

Mankiw, N.G., J.J. Rotemberg and L.H. Summers (1985). "Intertemporal substitution in Macroeconomics" *Quarterly Journal of Economics* **100**(1), 225-251.

Manser, M.E., and R. J. McDonald (1988). "An analysis of substitution bias in measuring inflation, 1959-1985", *Econometrica* **56**, 909-30.

References (continued)

Mattei, A. (2000). "Full-scale real tests of consumer behavior using experimental data", *Journal of Economic Behavior and Organization* **43**, 487-497.

Moran, L.R. and C.P. McCully (2001). "Trends in Consumer Spending, 1959–2000" *Survey of Current Business* **80**(3), 15-21.

Nelson, C.R. and C.I. Plosser (1982). "Trends and random walks in macroeconomic time series: some evidence and implications", *Journal of Monetary Economics* **10**, 139-162.

Nordhaus, W.D. and J. Tobin (1972). "Is growth obsolete?"in: *Economic Growth, Fiftieth Anniversary Colloquium*. Vol. 5. New York: NBER

Nordhaus, W.D. and J. Tobin (1973). "Is growth obsolete?" in: *The measurement of economic and social lPerformance*. M. Moss (ed.). Studies in Income and Wealth, Vol.38, 509-532. New York: Columbia University Press.

Patterson, K.D. (1991). "A non-parametric analysis of personal section decisions on consumption, liquid assets and leisure" *Economic Journal* **101**(408), 1103-1116.

Pollak, R.A. (1971). "Conditional demand functions and the implications of separable utility" Southern Economic Journal **37**(4), 423-433.

Robinson J.P. and G. Godbey (1997). *Time for life: the surprising ways Americans use their time*. University Park, PA: The Pennsylvania State University Press.

Rossana R.and J.J. Seater(1995). "Temporal Aggregation and Economic Time Series," *Journal of Business & Economic Statistics* **13**, No.4 (October), 441-451.

Rossiter, R.D. (2000). "Fisher ideal indices in the National Income and Product Accounts", *Journal of Economic Education* (Fall), 363-373.

References (continued)

Ruggles N.D. and R. Ruggles (1970). *The design of economic accounts*. New York: Columbia University Press.

Ruggles, R. and N.D. Ruggles (1982). "Integrated economic accounts for the United States, 1947-1980" Survey of Current Business **62**(5), 1-53.

Sippel, R. (1996). "A note on the power of revealed preference tests with Afriat inefficiency" *Projektbereich Discussion Paper* A-528.

Stock, J.H. and J.H. Wright (2000). "GMM with weak identification", *Econometrica* **68**(5), 1055-1096.

Swofford, J. L., and G.A. Whitney (1987). "Non-parametric tests of utility maximization and weak separability for consumption, leisure, and money", *Review of Economics and Statistics* **69**, 458-464.

_____(1988). "A comparison of non-parametric tests of weak separability for annually and quarterly data on consumption, leisure and money" *Journal of Business and Economic Statistics* **6**, 241-246.

_____ (1994). "A revealed preference test for weakly separable utility maximization with incomplete adjustment" *Journal of Econometrics* **60**, 235-249.

Triplett, J.E. (1992). "Economic theory and BEA's alternative quantity and price indices." *Survey of Current Business* (4). Washington, D.C.: U.S. Department of Commerce.

Varian, H. (1982). "The nonparametric approach to demand analysis", *Econometrica* **50**, 945-973.

_____(1983). "Nonparametric tests of consumer behavior", *Review of Economic Studies* **50**, 99-110.

References (continued)

_____(1988). "Revealed preference with a subset of goods", *Journal of Economic Theory* **46**, 179-185.

_____(1990). "Goodness of fit for revealed preference tests", *Journal of Econometrics* **46**, 125-140.

_____(1991). "Goodness of fit for revealed preference tests", *CREST Working Paper* 13(September), University of Michigan.

_____(1996). "Efficiency in production and consumption", in: H. Varian, ed., *Computational Economics and Finance*, New York: Springer-Verlag.

Wilcox, D.W. (1992). "The construction of U.S. consumption data: some facts and their implications for empirical work", *American Economic Review* **82**(4), 922-941.

Zolotas, X. (1981). *Economic Growth and Declining Social Welfare*. Athens: Bank of Greece.

Appendix on Data Sources and Methods

As mentioned in the main text, the datasets used here are primarily extracted from the BEA's NIPA and Fixed Asset tables – personal consumption expenditures and the depreciation/stock of consumer durables, respectively. Data sources and manipulations are presented in four parts, describing annual and quarterly data built under both new and old methodologies. The whole dataset is available upon requests in a spreadsheet (alldata.xls), and its notation is also described next.

A1.1-Annual Data under New Methodology (1929-2000)

The proxy for quantities are per capita real expenditures in 1996 chained dollars, whereas prices are Fisher indices (chain-type) normalized to 1996=100. Per capita figures are calculated using U.S. total population (Census Bureau), which excludes Hawaii and Alaska on years previous to 1950. The general notation is as follows:

d1, d2, d3: Per capita stocks of 1)Motor vehicles & parts, 2)Furniture & houshd equip., 3)Other durables.

nd1,..., nd5: Per capita real expenditures in 1)Food, 2)Clothing & shoes, 3)Gasoline & Oil, 4) Fuel Oil & coal and 5)Other nondurables.

s1, ..., s6: Per capita real expenditures in 1)Housing, 2)Household operations, 3)Transportation, 4)Medical care, 5)Recreation and 6)Other services.

ucd1,...ucd3: User cost of durables are calculated as described in the main text, using 6-month commercial paper rate (secondary market) and NIPA price indexes of new durables; as for the depreciation rates, BEA's Fixed Asset tables actually provide data on the annual flow of depreciation and on end-of-period net stocks of different categories of consumer durables. Annual depreciation rates can be calculated from the relationship between gross and net stocks of capital: $[K^{NET}=(1-\delta).K^{GROSS}]$. Note that, whenever used, "pf" stands for perfect foresight, in contrast to the other benchmark expectation model "stat", for the calculation of user costs under static expectations. A couple of (abnormal) negative estimates for the user costs of durables found over war years – early 40's, not consecutive observations – were linearly interpolated, as the GARP test requires positive prices and quantities.

pl and l: Leisure price/quantities. We started calculating current value expenditure in leisure, assuming 10-hour fixed allocation of time for sleeping and eating; from 7x14 hours of free allocation of time over a week, we subtracted the number of hours worked (see additional comments below), and multiplied the result by 52 to get an annualized estimate of hours allocated to leisure. Total expenditure at current values was obtained by multiplying the previous estimate by average hourly wage for each period, all data from BLS. We finally converted the series on hourly wages into a price index 1996=100 and used this index to calculate real expenditure on leisure time in 1996 dollars.

A1.2-Annual Data under Old Methodology (1929-2000)

Quantities in this case are in 1987 fixed-weight dollars, prices are proxied by Laspeyres indexes (fixed-weight) normalized to 1987=100. All data sources and methods are precisely the same as before, except:

(i) Before revisions over the 90's, NIPA reported only 5 subcategories of services, so that recreation was included in s5 (other services);

(ii) Depreciation rates were calculated as the reciprocal of numbers of service years that a durable good was expected to provide; as FHS actually worked with a larger number of subcategories (different aggregation level), we calculated the number of service years of durables in a category as the weighted average of service years of its components.

(iii) Figures were all calculated from fixed-weight quantity and price indexes reported in Nipa and Fixed Assets Tables, not available online but published in various issues of the *Survey of Current Business* and *NIPA*.

A1.3-Quarterly Data under New Methodology

All data sources and methods are precisely the same as before, except from the following cases: first, as quarterly figures on stocks of durables are not available, we essentially interpolated end-of-period annual stocks in accordance to the usual equation for perpetual inventory [$k_t = i_t + (1-\delta)k_{t-1}$], following FHS and Campbell and Mankiw (1989). Because only annual investment and depreciation data are available, we used seasonally adjusted quarterly consumption expenditure on durables as a proxy for the acquisition of new capital (i_t) and fit fixed depreciation rates for the quarters of a given year, so that reported and calculated end-of-period stocks were the same (each year). The final quarterly stock series were obtained from the average of end-of-quarter figures. At the end of the sample (2000:IV), the calculated quarterly depreciation rates for d1, d2 and d3 were approximately 9%, 5% and 4.5%, respectively. Specially concerning the

depreciation rates of autos, trucks and other motor vehicles (D1), this "fitted" depreciation rate may seem too high. However, as pointed out in FHS, annual stocks are net of discards; besides, a faster depreciation of capital at early stages of its service life is consistent with the new methodology of NIPA data, concerning the geometrical pattern of depreciation rates over time. Notice that the same depreciation rates were later adopted in the calculation of quarterly user costs, for consistency.

Second, in the calculation of quarterly user cost of durables, we used 3-month treasure bill rates (secondary markets) as the relevant interest rate, as published by the Federal Reserve Bank at St. Louis' website (FRED II). To convert annual rates into quarterly ones we applied the usual formula: $(1+r_t) = (1+R_t)^{0.25}$

A1.4-Quarterly Data under Old Methodology

All data sources and methods are precisely the same as in A1.3, except that figures were calculated from fixed-weight 1987 indexes, which are not available online and had to be typed in from various issues of BEA's *SCB* and *NIPA*.