On the existence of well-behaved macro utility functions:

Reassessing the power of Varian's Revealed Preference test in consumption aggregates

Luiz Maia Filho[†]
(luiz@dlch.ufrpe.br)
Professor Adjunto, Depto. de Letras e Ciências Humanas, Universidade Federal Rural de Pernambuco, (UFRPE). Rua Dom Manoel de Medeiros, s/n. Dois Irmãos. Cep: 52171-900, Recife/PE. F:(81) 3320-6462

Resumo:

Pesquisas empíricas baseadas em modelos de agente representativo assumem que os dados de certas categorias de consumo agregado são consistentes com uma função utilidade estável e bem comportada. Tal hipótese – quase sempre presumida verdadeira, sem verificação – pode, todavia, ser testada. Este artigo se dedica a investigar o poder do teste do Axioma Geral de Preferência Revelada (GARP), desenvolvido por Hal Varian, naquela que parece ser a base de dados mais comumente presumida consistente com funções de utilidade bem comportadas em toda a pesquisa empírica: os gastos de consumo não-durável americano no período do pós-guerra. Uma análise detalhada dos métodos previamente empregados para simular comportamento irracional – e, conseqüentemente, para se determinar o poder do teste numa determinada base de dados – indica que eles falham sob condições nada atípicas. Assim, os algoritmos existentes são reconsiderados e modificados para se evitar conclusões tendenciosas sobre o poder do teste. Os resultados sugerem que o algorítmo modificado é menos suscetível a gerar estimativas viesadas do poder do teste nos agregados tipicamente utilizados, mas documentam uma vez mais o baixo poder do teste em dados anuais, relativamente ao observado em séries trimestrais.

Abstract:

Empirical research based on representative-agent models often assumes that subsets of aggregate consumption data are consistent with some stable, well-behaved utility function. Such strong and mostly "maintained" assumption can, nevertheless, be put to test. This paper investigates the power of Hal Varian's General Axiom of Revealed Preference (GARP) Test in what seems to be the dataset most widely admitted consistent with simple utility functions in all empirical consumption research: the U.S. postwar personal consumption expenditures data on nondurables and services. A careful analysis of methods to simulate irrational behavior and, therefore, to determine the power of the test in a specific dataset indicates that they ultimately fail to do so under quite typical circumstances. Existing algorithm are reconsidered and subsequently modified to avoid biased conclusions about the power of the test, especially in low frequency data. Results suggest that our modified algorithm is indeed less likely to generate biased estimates of the power of the test in most of the previously studied subsets of data, but also confirm that the power of the GARP-test is significantly lower within annual series than in quarterly ones.

Palavras-Chave: Preferência Revelada, Testes Não-Paramétricos, Gastos de Consumo, Poder do Teste, Experimento Monte Carlo, Racionalidade, Função Utilidade.

Keywords: Revealed Preferences, Nonparametric Tests, Consumption Expenditures, Power of Tests, Monte Carlo Experiment, Rationality, Utility Function.

Classificação ANPEC:	Área 7 - Microeconomia, Métodos Quantitativos e Finanças
JEL Codes:	C14, C82, D12

[†]Thanks to Adrian Fleissig, Walter Thurman, John Seater, José Angelo Divino, Luiz Kehrle and anonymous participants of seminars at North Carolina State University, Universidade Católica de Brasília and Universidade Federal Rural de Pernambuco, for many helpful comments and suggestions.

1-Introduction

Aggregation both along time and across goods are commonly regarded as maintained assumptions in empirical studies of macroeconomic consumption. For more than twenty years[1], 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 (Choi et al., 2007). This paper revisits and builds upon methods previously applied to verify the power of the GARP test[2] in actual datasets – that is, the likelihood of committing a "type II" error and accept the utility maximizing hypothesis when it is false.

As GARP-consistency and weak separability tests are run on presumably separable subcategories of consumption expenditures, the existence of a well-behaved[3] macro utility function rationalizing those figures may be rejected (Varian, 1982). But if a set of aggregates passes both tests, empirical researchers can justifiably set out maximization models for those goods – conditional on total expenditure in those categories – and save degrees of freedom in the estimation of preference parameters from usual optimality conditions (Swofford and Whitney, 1987; 1994; Drake, 1997; Fleissig, Hall and Seater, 2000; Andreoni and Miller, 2002; Blundell, Browning and Crawford, 2003).

Due to the nonparametric nature of the GARP test, researchers must be aware that two extreme outcomes may occur: first, the test can reject the existence of a well-behaved representative utility function rationalizing long series of data (H0) due to a single deviation from the expected maximizing behavior – which suggests that significant measurement errors could frequently lead to rejection of the null hypothesis (Jones and Peretti, 2005; Elger and Jones, 2008). Secondly, the test is likely to have low power in datasets where budget hyperplane intersections are infrequent. This second possibility is the one of most concern here, for it can arise in a fairly common circumstance: the use of low frequency data (annual figures) in periods of economic expansion. Therefore, temporal aggregation plays a role in one's assessment of the validity of a particular aggregation across goods.

Bronars'(1987) Monte Carlo experiment generated random budget shares to find that the test was more likely to commit a type II error in observed annual-budget hyperplanes than in quarterly ones. Since then, different approaches to generate the alternative hypothesis (irrational behavior) were developed. In the next sections, we will review those procedures, propose a modified algorithm, and observe which ones are more likely to under/overestimate the power of the GARP-test in quarterly and annual series of U.S. consumption expenditure data. As we do so, we argue that some of the algorithms are especially susceptible to failure, i.e., generating data quite consistent with rational behavior.

The remainder of this paper is organized in four additional sections. First, we review the relevant tools of revealed preference analysis and the methods to assess the power of the GARP-test in actual datasets. In section 3 we provide a brief description of data sources and manipulations, before considering the evolution of budget shares along the studied sample. Next we simulate data with different algorithms and run the GARP-test in both actual and constructed series. Final remarks, overall conclusions and a direction for future research are presented in the last section (5).

---

[1]The whole literature on Consumption-Based Asset Pricing Models, for example, have traditionally assumed – without testing – the existence of well-behaved utility functions rationalizing choices on either nondurable goods alone or the sum of nondurable goods and services, in investigations of intertemporal substitution and risk aversion; see Hansen & Singleton (1982), Stock & Wright (2000). See also Choi et al. (2007) for a new "toolkit" to the application of Varian's method.

[2]GARP stands for General Axiom of Revealed Preference; concepts and tests are discussed in the next section of the paper.

[3]For "well-behaved" utility function we mean one that is monotonically increasing, semi-concave and twice-differentiable.

2-Varian's tools of revealed preference analysis and their power

2.1-Garp and weak separability tests: a brief review

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):

$$x^t R^D x^s \quad <=> \quad p^t x^t \geq p^t x^s$$

Intuitively, we cannot say that $\mathbf{x^t}$ is revealed-preferred to any other bundle that was not 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, $\mathbf{x^t R x^z}$ if and only if there is some chain of observations ($\mathbf{x^t}$, $\mathbf{x^u}$, $\mathbf{x^v}$,…, $\mathbf{x^z}$) such that $\mathbf{x^t R^D x^u}$, $\mathbf{x^u R^D x^v}$,…, $\mathbf{x^y R^D x^z}$.

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. Thus, if $\mathbf{x^s}$ and $\mathbf{x^t}$ were affordable at both occasions, the consumer should have made the same pick. 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:

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 time.

As for the nonparametric test of weak separability, which requires GARP-consistency as a necessary condition, it was also originally developed and implemented by Varian (1983). As before, one can avoid restating theorems and proofs, but it is useful to define and illustrate weakly separable/nonseparable preferences before we further discuss the test and distinguish between necessary and sufficient conditions of the test.

Following Fleissig, Hall and Seater (2000), suppose that a vector of k goods is partitioned into two subsets, **a** and **b**, where $\mathbf{a}=(x_1, x_{2,\ldots}, x_m)$ and $\mathbf{b}=(x_{m+1}, x_{m+2}\ldots, x_k)$; a utility function $U(\mathbf{x})$ is weakly separable in b-goods if there exist a subutility function $v(\mathbf{b})$ and a macro function $u^*[\mathbf{a}, v(\mathbf{b})]$ 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[4].

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, we can denote $U_i(x)$ 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 words, the marginal rate 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 pass the GARP-consistency test, due to the fact that it could reject the existence of a well-behaved subutility function v(.). Sufficiency is achieved if there are numbers satisfying a series of inequalities – known as Afriat inequalities – involving prices, quantities and expenditures on each supposedly separable subset of goods (Varian, 1983). Fleissig and Whitney (2003) developed a relatively efficient algorithm to search for those 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)[5].

In practice, 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

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

[5] The majority of 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.

convenient, but actually a theoretically valid procedure. Passing none of them, on the other hand, leads to a conclusive rejection of some separability structure. However, passing the necessary condition only means that the algorithm used to search for numbers satisfying the aforementioned inequalities failed to find them. The result is inconclusive because the existence of such numbers is not definitely rejected.

2.2-Bronars'(1987) approach and the power of the test

Bronars' simulation approach was proposed to address particularly 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, evidence suggests that the nonrejection of the actual choices was not 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[6].

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[7]:

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

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.

[6] Also central to Bronars' discussion was 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.

[7] 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.

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 is not necessarily true, especially if the actual budget lines shift significantly over time; if budget lines do not intersect in the actual dataset, they will not intersect in randomly simulated series. GARP violations will never occur in either case.

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:

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

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. 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.

2.3-Extensions of the original 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 rational and irrational choices, all leading to the same observed expenditure. That is also the source of its most notorious shortcoming: conclusions from the application of such method are sensitive to the particulars of the budget-share simulating algorithm.

Unless the test is shown to have high power in a dataset with the adoption of multiple irrationality models, it may be hard to establish how truly strong is the evidence of utility-maximizing behavior coming from the GARP-consistency of a dataset. Let's consider a couple of alternative algorithms proposed in the literature as extensions to Bronars'. Later, we will suggest a new one, which may have advantages in replicating relevant characteristics of actual data.

Recall 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[8], the allocation of resources on consumption subcategories are expected to change over time, not returning to an overall average value.

[8]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.

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 choices are (likely) similar to those for weakly separable goods derived from a Cobb-Douglas utility function[9].

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.

Generating random but GARP-consistent choices and, for that reason, underestimating the power of the test in a dataset is also possible – and for the same reasons mentioned above – if one adopts two other proposed algorithms, that we discuss next.

Burton (1994) applied and extended the original 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:

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

$$S_{nt\,(3)} = 1 - \sum_{i=1}^{n-1} W_{it} \tag{4}$$

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 random choices that involve budget share allocations very similar to the observed ones.

Cox (1997) proposed another alternative hypothesis to introduce "irrationality" through randomness but preserving characteristics of real data. He opted for 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

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

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

Cox himself observed that the GARP test would likely fail to reject his random (simulated) data 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. As before, it would lead to biased conclusions about the power of the test. Concerns of this nature suggest that the actual evolution of budget shares should be considered previously to the application of method of power assessment. It also motivates the construction of an algorithm that incorporates observed trends in budget shares, as we do next.

2.4-A new simulation algorithm: incorporating consumption trends

A straightforward adjustment to Bronars' second algorithm [$S_{it}(2)$] permits incorporating consumption trends into the simulated series. It simply involves replacing the historical average budget shares across all years (K_i) in expression (2) with moving averages (L_{it}) of observed budget shares across some range of periods around each time t; the expression would then become:

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

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.

This 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 historical average as the parameter from which simulated budget shares would randomly deviate at each period, we arbitrarily pick the mean budget share of, say, 5 periods over the local interval [t-2,t+2]. Four observations end up being lost (at the beginning and in the end of the simulated series), but we may gain in the sense that these simulated data do not deviate systematically from the actual evolution of budget shares over the sample.

3-The data

3.1-Main sources and methods

The two main sources of data used here are BEA's NIPA (*National Income and Product Accounts*) and *Fixed Assets* tables, for seasonally adjusted personal consumption expenditures and the depreciation / stock of consumer durables, respectively. Both quarterly and annual data were collected for the period from 1964 to 2000[10].

[10] The most relevant limitations came from labor data, with required series being available starting in 1964. Therefore, whenever we run test on datasets that do not consider leisure choices, the actual sample period was 1959-2000.

Following once more Fleissig, Hall and Seater (2000), we adopted not the major categories of real personal expenditures but their components: a set of 14 subcategories of consumption expenditures, besides leisure:

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

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.

As for leisure prices and quantities, we essentially followed the procedures in Swofford and Whitney (1987) and assumed a 10-hour daily fixed allocation of time for sleeping and eating. 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. Leisure time (quantity) is residual, as one subtracts the average number of hours actually worked from the daily 14 hours available. The opportunity cost of time is *proxied* by the wage rate – seasonally-adjusted average hourly earnings of nonsupervisory workers on private nonfarm payrolls. All labor data provided by the Bureau of Labor Statistics[11].

Also in the case of durables we adopted standard procedures in this literature: consumers are assumed to obtain utility from services that are 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[12]:

$$uc_t = p_t - [(1-\delta_t)/(1+R_t)] \cdot E_t(p_{t+1}) \quad \textbf{(6)}$$

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. Two benchmark expectation models were assumed: static expectations – the expected price one period ahead is simply today's figure – and perfect foresight[13].

Worth mentioning, we deliberately avoided the use of GARP and weak separability tests on monthly data because those are particularly subject to measurement problems; Wilcox(1992) pointed out two

[11] 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."

[12] Both Swofford and Whitney (1987, 1988) and Fleissig, Hall and Seater (2000) 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 (6).

[13] Results were not sensitive to the choice of expectations models. We report results obtained from the dataset calculated assuming perfect foresight of the future price of durables held by the consumer. A spreadsheet with all series of prices and quantities used in this paper will be promptly made available upon request to the author.

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. His main conclusion was that published data – especially monthly figures – cannot always be assumed to correspond exactly to their theoretical analogues[14].

3.2-Evolution of budget shares

Our discussion in the previous section 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. Upon inspection, one can verify that many consumption subcategories seem to wander inside relatively narrow ranges of budget share allocations (ND4, ND5, S1, S2, S3), but others such as Food (ND1) and Medical Care (S4) have clearly shifted over those years. In the case of Medical Care, its share rose steadily from approximately 1.5% of total expenditure to about 5.5% at the end of our sample.

4-Data simulation and test results

We must begin this assessment of the power of the GARP-test by stating that the actual series, at both frequencies, passed the test; that is, the original data contained no violation of GARP. Therefore, we can proceed and study how confident on such result one can be. Let's compare – firstly – results from simulated data at both frequencies, with and without leisure (as one of the relevant consumption subcategories). In table 1, Q1 and A1 refer, respectively, to quarterly and annual datasets including all subcategories of durables, nondurables, services and leisure. The last two datasets at each frequency exclude leisure (Q2, A2). In each case, we generated 2000 simulated series using Bronars' first method [$S_{it}(1)$] and the same number of series with his second algorithm [$S_{it}(2)$] – for starters.

An 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[15]. Suppose, for example, that an exogenous and large 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, fluctuations of the price of leisure can also increase the number of budget intersections, using the same logic discussed above, provided that shifts are not always positive or negative. The issue can be empirically investigated with simple computations of how often budget hyperplanes do intersect within both original datasets, as we also report in table 1.

[14] Fleissig, Hall and Seater 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).

[15] 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. Nevertheless, 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.

Table 1 - 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.

As we inspect first the results for quarterly data (Q1 and Q2), 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 "line" intersects with others is very large, around 60% of the possible number of times. Those results seem to corroborate the high power of the test over datasets with often intersecting hyperplanes.

The GARP-test results with annual data indicate that its power 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 discussed before.

Next, table 2 reports estimates of the power of the test using all simulation algorithms considered previously (including once more the number of budget intersections observed in each dataset). The use of all five methods confirmed, to a large extent, the overall conclusions drawn from table 1: the power of the test is indeed very high within quarterly datasets, regardless of leisure inclusion or simulation methods. However, the test is revealed to be substantially weaker at the annual frequency against the majority of alternative hypotheses – than against Bronars'.

Table 2 - 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.

Randomly simulated quarterly datasets were rejected quite often, at least 96% of the times. Just like with Bronars' simulation methods [$S_{it}(1)$, $S_{it}(2)$], all 2000 quarterly datasets were rejected with the new algorithm [$S_{it}(5)$], 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: only in those cases, the estimated power of the test fell when leisure was included. 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 from 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 times 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 (see table 1). The explanation for this puzzler issue is that randomness does not necessarily imply irrationality.

The common characteristic of those two simulation methods [$S_{it}(3)$, $S_{it}(4)$] 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 from the actual (possibly rational) ones.

Since actual budget shares for many of those goods lie along relatively narrow ranges of values, the chances of simulating data that are significantly different from the actual 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 $S_{it}(4)$: the inclusion of leisure involves implicitly assuming that an outstandingly large budget share is allocated to a single good at all times. It necessarily makes narrower the observed ranges of budget shares of other goods, as they become relatively less significant. 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 the puzzle is the referred simulation methods have merely failed to generate irrational data.

As for the issue of the estimated power of the test being so lower with Bronars' second algorithm than with his first, apparently $S_{it}(2)$ generates data that are, on average, consistent with GARP. Simulated budget shares for each good "fluctuate" around their historical average, equivalently to the ones possibly derived from a Cobb-Douglas utility function with some sort of random errors.

The new simulation method seem to have 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 rational consumer. $S_{it}(5)$ seems to be a reasonable alternative for it preserves both 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 – an the eventual long-run trends of actual consumption data, which were not incorporated within Bronars' methods.

5-Final remarks and conclusions

In this paper we discussed circumstances under which Bronars' (1987) simulation algorithms may lead to biased conclusions about the power of the GARP-test, which verifies the existence of well-behaved utility functions rationalizing particular sets of personal consumption data. The original algorithm and some alternative ones generate random data that are either too extremely unreasonable (overestimating the power of the test) or much likely GARP-consistent ones (underestimating the power of the test). A trivial adjustment to Bronars' original algorithm was proposed and applied, so as to incorporate information about the evolution of observed budget shares along the sample and, therefore, reduce the likelihood of the mentioned biases. More extensive use of Monte Carlo experiments seems to be a promising next step, in the attempt to establish the superiority of any of those algorithms, though – a direction for future research.

We also confirmed 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.

The low power of the test within annual datasets is an important result by itself, but it also provides a simpler explanation than Swofford and Whitney's (1987, 1988) for the nonrejection of GARP-consistency on annual datasets that included broad monetary aggregates (Money-in-Utility Function models); they argued 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. That is, indeed, possible, but our new explanation based on the power of the GARP test has the advantage of not relying on unobservable factors (adjustment costs).

References

Andreoni, J. and J. Miller (2002). "Giving according to GARP: An experimental test of the consistency of preferences for altruism". *Econometrica* **70(2)**, 737-753.

Blundell, R.W., Browning, M. and I. Crawford (2003). "Nonparametric Engel Curves and Revealed Preference" *Econometrica* **71(1)**, 205-240.

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.

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

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.

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

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

Choi, S., Fisman, R., Gale, D.M. and S. Kariv (2007). "Revealing Preferences Graphically: An old method gets a new tool kit" *American Economic Review* **97(2)**, 153-158.

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

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

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.

Elger, T. and B.E. Jones (2008). "Can rejections of weak separability be attributed to random measurement errors in the data?" *Economic Letters* **99**(1), 44-47.

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.

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

References

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.

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

Jones, B.E. and P.D. Peretti (2005). "A Comparison of two methods for testing the utility maximization hypothesis when quantity data are measured with error" *Macroeconomic Dynamics* **9**, 612-629.

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

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

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

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.

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.

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

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.