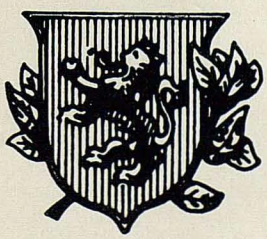
REVISTA DE LA ACADEMIA DE CIENCIAS

Exactas
Físicas
Químicas y
Naturales

DE ZARAGOZA



Serie 2.ª
Volumen 49

1994

INDICE DE MATERIAS

Rev. Academia de Ciencias. Zaragoza. 49 (1994)

ANTINOMICITY AND THE AXIOM OF CHOICE: A CHAPTER IN ANTINOMIC MATHEMATICS

To the memory of Stanisław Jaśkowski

F.G. ASENJO
Department of Mathematics
University of Pittsburgh
Pittsburgh, PA 15260

I. Introduction and Motivation

§1. A positive view of antinomies.

Russell's discovery in 1902 of the antinomy of the set of all sets which are not members of themselves prompted a profound and widespread examination of the foundations of mathematics for many years to come.[1] No other discovery has shaken mathematics and logic more deeply than Russell's antinomy, which came just when set theory was beginning to be widely accepted after years of rejection. From that time on antinomies have been treated seriously — to be avoided, to be sure, but nevertheless constituting a stimulating logical phenomenon at the very heart of mathematical reasoning.

The next natural step was the acceptance of antinomies in their own right. For this to occur, the basic logical assumptions had to be changed; this was accomplished in various ingenious ways, setting aside in the process the many acrobatic pirouettes that logic had been required to perform in order to jump over antinomies without tripping.

Underlying this acceptance is the belief that there is something intrinsically valuable in antinomies. Evolving from being merely a strong motivating force for deep analysis of the foundations of mathematics, antinomies now became a significant center of attention in themselves, a positive part of reason with their own legitimacy. This legitimacy arises from the fact that, although not always so, our thought processes are often antinomic, which in turn reflects the parallel fact that reality itself is often antinomic — hence why not logic and mathematics?

What began as a few timid investigations today has proliferated into a vast variety of logical approaches, different in point of view and method but all sharing in common the

objective of using antinomies positively as valuable, intelligible, and rational parts of the logical discourse.[2]

The many antinomic logics now in existence prove beyond question the feasibility of the formal incorporation of antinomicity as an extension of rationality. What is still missing, though, are the strictly mathematical applications of this logical approach. In order to obtain acceptance of antinomic logic as more than a curiosity, new and effective mathematical structures must be developed — as happened with nonstandard models, in the limbo of curiosa before A. Robinson put them to good use. The present work is an attempt to break ground in mathematics proper, armed with the accepting view just described. Specifically, we shall examine various versions of antinomic set theory, in particular the axiom of choice, keeping the presentation as intuitive as possible, more in the manner of a nineteenth century paper than as a thoroughly formalized system. The reason for such a presentation is the conviction that at this point it should be the mathematics that eventually determines the logic, rather than the other way around.

§2. Some antecedents of this view.

Kant was the first modern thinker to make the point that antinomies are not to be "solved" but accepted as constructive rational elements. In his *Critique of Pure Reason* he presents them not only as a reflection of the nature of the mind but also as a force to awake reason from its consuetudinal state of slumber.

Cantor was the first mathematician to acknowledge the presence of inconsistencies in set theory but he left them alone, only mentioning them casually in a letter published for the first time in 1932. He said: "For a multiplicity can be such that the assumption that *all* of its elements 'are together' leads to a contradiction, so that it is impossible to conceive of the multiplicity as a unity, as 'one finished thing.' Such multiplicities I call *absolutely infinite* or *inconsistent multiplicities*."[3] Also, "Two equivalent multiplicities either are both 'sets' or are both inconsistent."[3] Further, Cantor was not particularly upset by Russell's discovery (as Frege was), having himself discovered in 1895 the "paradox of the largest cardinal number."

Although a Platonist and therefore a believer in the reality of correct mathematical propositions, Gödel admitted "the amazing fact that our logical intuitions (i.e., intuitions concerning such notions as truth, concept, being, class, etc.) are self-contradictory."[4] He added that it is "not self-contradictory that a proper part should be identical (not merely equal) to the

whole, ... and it is easily seen that there exist also structures containing infinitely many different parts, each containing the whole structure as a part."[5] "Furthermore, there exist sentences referring to a totality of sentences to which they themselves belong."[6]

II. Logic for Antinomies

§3. Prelogical antinomies.

Before the antinomies of simultaneous truth and falsity, we have the antinomies of sense and nonsense — Zeno's paradox being one such prelogical antinomy. And, before the latter, we still have concrete apophantic antinomies, i.e., factual antinomies in which two opposite qualities are displayed simultaneously in the same given entity or event; for example, when something is both plural and unitary, or when a movement both helps and hinders reaching an objective, etc. However, even though an antinomic logic can be based on an antinomic semantics of sense and nonsense, so too can such semantics be based on the phenomenological apophantic description of a contradictory reality. We shall put aside these last two important prelogical areas, for they must be treated at length elsewhere. Here, we shall deal only with the true-and-false kind of antinomy, treating it as the absolute beginning. We must always keep in mind, however, that any talk of difference in identity — say, of something being simultaneously the same and different, which occurs naturally and correctly in ordinary language — already involves antinomic thinking and points implicitly to the coexistence of truth and falsity.

§ 4. Truth-and-falsity not a third logical value.

Sometimes truth is simple and so is falsity, but at other times we hit upon the true by way of the false in a way that makes the false a necessary component of the true. To see the true in the true-and-false as different from the true in truth alone is not an accurate conception of antinomicity. To fit the facts, the logic of antinomicity should not be conceived as a three-valued logic but as a complex two-valued one in which truth valuations are not functions but rather one-to-one or one-to-two correspondences between sentences and the unordered pair $\{T,F\}$. That is, some valuations assign to a sentence A the value T, to a sentence B the value F, and to a sentence C both values T and F.

§ 5. Assertion and negation independent of truth and falsity.

As mentioned, there are already many antinomic logics in the literature, with more to come. The logic outlined here is clearly not the only possible one. In accordance with the comment at the end of §1 to the effect that we do not want to reflect prejudice towards any of the new directions that will arise from mathematical applications, we shall set here only a minimum list of conditions that are indispensable for our own objectives, leaving the rest indefinite. It should be obvious, for example, that reduction to the absurd as a method of proof cannot be allowed if one is to avoid the derivability of every well-formed formula from a single contradiction, i.e., a fall into absolute inconsistency.

It is indeed extraordinary that even the most sophisticated definitions of truth and falsity, as well as those of assertion and negation, rely on one another in a blatant vicious circle. Thus, for example, the truth of an atomic predicate formula is defined in accordance with the interpreted terms of the formula belonging or *not* belonging to the set-theoretic relation that interprets the predicate, a metalinguistic definition that leans on negation and set theory as much as it does on the law of excluded middle — the n-tuple of terms $\langle t_1,...,t_n\rangle$ in $P(t_1,...,t_n)$ is a member of the relation R that interprets the n-ary predicate P, or it is not, one or the other, with no third alternative possible. In turn, the semantic definition of negation is given in terms of truth values as follows: If a sentence A is true, its negation $\neg A$ is false, and if A is false, then $\neg A$ is true. This definition is taken to be the last word on the matter, and is uncritically used whenever negation is used (except that within many-valued logics — which lie beyond our scope — a different approach to negation is considered). In the classical propositional calculus, then, given a sentence A and a structure $\mathcal{I}$ that interprets the language in which the sentence is formed, the sentence is either simply true or simply false; in symbols, $\models A$ or not-$\models A$, but not both. Further, syntactically only A or $\neg A$ are provable; again, no third alternative is allowed.

Here, however, negation will be looked at differently, accepting the principle that negation is not a logical operation definable in terms of truth and falsity, but that its meaning, in effect, stands prior and beyond whatever any truth table rule can provide. Russell has already observed that there is something primitive and peculiarly irreducible in the notion of negation that escapes the truth-table approach: he believed in the necessity of some "negative basic propositions" side by side with the positive ones[7] — fundamental propositions, atomic in their own way. This belief, followed systematically, sets negation apart from the other

connectives — which is precisely the objective of this section. Our reasoning, however, arises from the antinomic approach, that is, by the acceptance of the fact that there are cases in which one can see contradictions in a negation — that $\rceil A$ is true-and-false — and hence, that truth is not necessarily simply the negation of a falsity. Or even more strikingly, we come to the same reasoning by recognizing the existence of cases in which a negation is neither true nor false. Antinomies, which themselves do not necessarily depend on negation, force upon us the inevitable conclusion that truth and falsity must be divorced from assertion and negation, that $\rceil A$ may be simply true, simply false, true and false, or neither true nor false.

To make all this intuitive let us say informally that the negation of a sentence A refers to all assertions A_i that are in opposition or disagreement with A. Negation, therefore, is a form of indirect assertion; as such, it can be characterized as a mapping on the class S of all well-formed sentences of a given language $\mathcal{L}$ into the power set of S as follows: $\rceil A=\{A_i: i\in I\}$ where the indexed A_i's (a finite or infinite family) are all the assertions in $\mathcal{L}$ which stand in opposition to A. If, for example, we consider A to be the sentence "four is even," there is of course only one sentence in opposition to A: "four is odd," hence, the semantic meaning of the expression $\rceil A$ is in this case the singleton {"four is odd"}. In fuzzy set theory on the other hand, if A stands for $a\in_0 b$ (a is a member of b with probability zero), then $\rceil A$ is the uncountable set of sentences $\{a\in_r b$: where r is a real number such that $0<r\leq 1\}$. It has been suggested that $\rceil A$ means to assert the disjunction $A_1\vee A_2\vee ...\vee A_n$ of a finite sequence of sentences in opposition to A. Apart from the A_i's possibly being infinite in number as in the last example, this interpretation of negation would subject it to the truth-table definition of disjunction. Since we want to have negation fully independent of all other connectives, we shall adhere to the "neutral" set-theoretic characterization given above — not a definition proper but an informal intuitive one similar to Cantor's characterization of a set as "a multiplicity taken as a unit," (which, incidentally, was Kant's characterization of a totality).

As for the truth of a sentence A, we can simply say — also informally — that A is the case in a given context. Now, A can be fully the case — simply true — or only partly the case — true-and-false. Further, A is simply false if the context fully opposes A, and neither true nor false if the context is fully irrelevant to A. This is all we shall say here to make understandable our having four possible truth valuations (T, F, $T\&F$, neither T nor F) for a given sentence, plus four truth valuations for its negation. Note, again, that negation does not determine truth and falsity but is given either a single value, two values, or none, regardless

of the truth values for the corresponding assertion. In other words, each of the four cases for A branches out into four additional cases for $\neg A$: negation extends assertion, does not exclude it. Needless to say, we shall make room for the preservation of the standard two-value situation of ordinary sentences, i.e., "two is even" is to remain true and "two is not even" false, etc. Whether the expression above, "to be the case," is an assumption or is to be determined by an effective rule whenever possible depends on mathematical considerations that cannot be established in advance.

Now, a theory can be considered complete in two metalinguistic senses: (i) every sentence A is tautologically true or logically valid if and only if it is provable from the theory's axioms; in symbols, $\models A$ iff $\vdash A$ (completeness theorem) and (ii) given a sentence A in a theory $\mathcal{T}$, either A is a theorem or $\neg A$ is: $\vdash A$ or $\vdash \neg A$ (definition of complete theory). Classically, the metatheorem "(i) implies (ii)" is proved by metalinguistic contraposition and the law of excluded middle. That is, if (ii) fails there is a sentence A which is neither provable nor refutable, but A must be true or false according to the law of excluded middle that tells us $\models A$ or $\models \neg A$, which contradicts (i). The metatheorem "(ii) implies (i)" is proved in a similar way by contraposition and contradiction. In contrast, an antinomic logic sets aside proofs by contradiction and the laws of contraposition and excluded middle (although the latter will be acceptable for metalinguistic statements, as will be made clear below). Hence, (i) and (ii) are to be considered simply hypotheses independent of one another. All this means that while $\vdash A$ and $\vdash \neg A$ are both possible simultaneously, if $\models A$ is the case, not-$\models A$ cannot also be the case; although A can be true and false, not-$\models A$ (A is not true) is not synonymous with false. The metalinguistic contradiction $\models A$ and not-$\models A$ is not allowed: contradictions belong to the object language. Furthermore, (i) applies to true formulas but says nothing about false ones: a false formula may be provable or not, even if (i) is assumed. Also, the law of excluded middle does not extend to the metalinguistic statement "$\models A$ or $\models \neg A$," for neither $\models A$ nor $\models \neg A$ may be the case if both A and $\neg A$ are simply false, say; on the other hand, $\models A$ and $\models \neg A$ may also be compatible. Finally, not-$\models A$ and not-$\models \neg A$ may both be the case simultaneously (again, keep in mind that not-$\models A$ is not the same as A is false).

Metalinguistically, then, negation preserves some of the characteristics of its classical use; for example, as already indicated, no metalinguistic assertion or negation is both true and not true, and it must be either one or the other. The metalinguistic "not," then, abides by no-contradiction and excluded middle, in contrast to the object-language negation "$\neg$" which

will not satisfy either. However, the laws of contraposition and double negation (in both directions) and the proofs by contradiction will not be valid in either the metalanguage or the object-language of our antinomic logic. These proof-theoretic limitations have been adopted to broaden negation's meaning and are not more restrictive than those of intuitionism, which rejects the laws and proof method just mentioned, except for $A \Rightarrow \neg\neg A$ and $(A \Rightarrow B) \Rightarrow (\neg B \Rightarrow \neg A)$. In addition, antinomic logics are for the most part nonconstructive and are therefore in a stronger proof-theoretic position than intuitionism to find deductive replacements for the laws and proof method in question. For example, we shall assume completeness in the sense of (i), i.e., $\models A$ iff $\vdash A$; thus, the proof of the semantic truth of A will automatically entail A's syntactic provability. In addition, the systems to be proposed can be extended to complete extensions in the sense that for every sentence A either $\vdash A$, or $\vdash \neg A$, or both, extensions in which because of (i), either $\models A$, or $\models \neg A$, or both must be the case respectively. (Once more, note that "$\vdash A$ and $\vdash \neg A$" does not imply "$\models A$ and not-$\models A$" but merely "$\models A$ and $\models \neg A$" as stated.)

For A false let us use the symbol $\nvDash A$, given that falsity, just as negation, will have a positive meaning here, not a negative one: A false may mean A holds in a different context, or in the same context relative to a different rule than the one that makes A hold, if the latter is indeed the case. This is patently clear in model theory, where the truth of a sentence is a function of the domain of interpretation (the universe of discourse or context of the moment) and the specific rules attached to that interpretation. This relativity of truth and falsity will be expanded: not only will there be no absolute truth and no absolute falsity but also truth and falsity will not be truth values rigidly connected to one another.

Metamathematically, then, although simultaneously we can have $\models A$ and $\nvDash A$, we cannot have $\nvDash A$ and not-$\nvDash A$, and at most we can have one or the other; nor can we have $\nvDash \neg A$ and not-$\nvDash \neg A$, but at most only one or the other. The metalinguistic rule of excluded middle which applies to "not" does not, however, extend to the following: $\models A$ or $\nvDash A$, $\nvDash A$ or $\nvDash \neg A$ (A and $\neg A$ may both be not false), not-$\models A$ or not-$\nvDash A$ (A may be true and false). The following are possible though: (i) $\nvDash A$ and $\nvDash \neg A$, (ii) not-$\models A$ and $\nvDash \neg A$, and (iii) $\models A$ and not-$\nvDash \neg A$. Similarly, we cannot have $\vdash A$ and not-$\vdash A$, or $\vdash \neg A$ and not-$\vdash \neg A$. Nor is it the case that if $\vdash A$ then not-$\vdash \neg A$, or that if $\vdash \neg A$, then not-$\vdash A$. (Incidentally, were we to have several truth values $t_1, t_2, \ldots, t_m$, and several false values $f_1, f_2, \ldots, f_n$, using them to distinguish $\models_{t1} A, \ldots, \models_{tm} A$, $\nvDash_{f1} A, \ldots, \nvDash_{fm} A$, the metalinguistic law of excluded middle would extend in the sense that $\models_{ti} A$ or not-$\models_{ti} A$ but not both, and either $\nvDash_{fi} A$ or not-$\nvDash_{fi} A$ but not both).

We shall emphasize that, although a sentence in a given language is designated as true or false, or both, or neither, in accordance with context and interpretation, these designations need not be understood in set-theoretic terms. An explicit assumption, or a constructive or nonconstructive rule is indispensable of course, but assumptions and rules can be presented in many forms that are decidedly independent of set theory. Also, whereas in the classical propositional calculus the class of all the negations of tautologies is a disjoint mirror image of the class of tautologies, here — with the broader meaning of negation — the two classes intersect and in the class of tautologies we shall find both propositions and their negations.

Finally, antinomic logic makes room for an included middle, which intuitionism will abhor. In antinomic logic if A and $\neg A$ are both simply false, and $A \vee \neg A$ is also simply false, the latter may not exclude a *tertium datur*, say $\neg\neg A$, which is a consequence of the fact that true and not-true, false and not-false, are metalinguistic assessments — and compatible ones at that. In other words, $A \vee \neg A$ is neither a tautology nor a contradictory statement (always false). It is a contingent statement, true or false as the case may be. Metamathematically, even both not-$\vdash A \vee \neg A$ and not-$\vdash \neg(A \vee \neg A)$ are contingent.

§6. Truth tables for the positive fragment of logic, and other assumptions.

Having made the point that antinomicity is better off with a revised negation that makes it a nonexclusive operation independent of any truth table, we must now turn to the remaining connectives of the propositional calculus — the positive fragment — as well as to the definitions of truth, of an antinomic model for the predicate calculus, and of first-order theories. We shall adopt the following tables for the four positive connectives.

$A \wedge B$

A\ \ B\	T	F	T&F
T	T	F	T&F
F	F	F	F
T&F	T&F	F	T&F

$A \vee B$

A\ \ B\	T	F	T&F
T	T	T	T
F	T	F	T&F
T&F	T	T&F	T&F

$A \Rightarrow B$

A\ \ B\	T	F	T&F
T	T	F	T&F
F	T	T	T
T&F	T	T&F	T&F

$A \Leftrightarrow B$

A\ \ B\	T	F	T&F
T	T	F	T&F
F	F	T	T&F
T&F	T&F	T&F	T&F

The following example shows how these tables were generated for the antinomic cases. If classically A is either true or false and B is true, then the compound statement $A \Rightarrow B$ is true for both cases; hence, if A is antinomic ($T\&F$) and B true, $A \Rightarrow B$ is true. If A is antinomic and B is false, $A \Rightarrow B$ is antinomic since it is false if A is true, and true if A is false. If A is false,

the truth value of B is irrelevant, $A \Rightarrow B$ is therefore true; hence, if A is false and B antinomic, $A \Rightarrow B$ is true. These tables are the same as those proposed in a previous paper.[8]

As for the syntax, we shall keep the two positive propositional axiom schemes given in Mendelson[9]: (i) $A \Rightarrow (B \Rightarrow A)$ and (ii) $(A \Rightarrow (B \Rightarrow C)) \Rightarrow ((A \Rightarrow B) \Rightarrow (A \Rightarrow C))$, dropping the axiom scheme (iii) $(\neg B \Rightarrow \neg A) \Rightarrow ((\neg B \Rightarrow A) \Rightarrow B)$, the last being the only propositional axiom scheme involving negation. We shall also keep $A, A \Rightarrow B \vdash B$ (modus ponens), applicable exclusively to positive statements, that is, statements in which "$\neg$" does not occur at all.

Let us call positive tautology any positive propositional statement that is either true or antinomic by construction as determined by the above truth tables. In turn, let us call negative tautology any propositional statement that involves at least one occurrence of negation and that is either true or antinomic by specific designation, whatever the truth values of all the positive statements involved. Schemes (i) and (ii) and modus ponens generate positive tautologies only, i.e., $\vdash A$ implies $\models A$, understanding $\models A$ to mean A is true or true-and-false.

We shall assume the completeness theorem as a meta-axiom for the propositional calculus, i.e., we now add $\vdash A$ if and only if $\models A$ for all well-formed statements, positive or negative. As a consequence, a negative statement automatically becomes a syntactic axiom whenever it is declared true or antinomic by specific designation, i.e., by an ad hoc assumption or rule, since no truth table or general axiom scheme regulates negation. In this manner, we are able to move freely not only from syntax to semantics but also from semantics to syntax.

For the predicate calculus, the usual notion of interpretation is to be expanded as follows. Given a domain of interpretation or universe D (a set-theoretic particularization of the broader concept of context), each predicate P of a formal language $\mathcal{L}$ is associated not only with one but with four relations R_1, R_2, R_3, R_4 such that if P is an n-ary predicate, the relations R_i (i=1,2,3,4) are all n-ary, and each is a subset of the Cartesian product D^n. In addition, formal terms t_1, t_2,... are interpreted in the domain D by specific individuals of D in the usual way. The interpreted terms will be denoted by $\mathbf{t}_1, \mathbf{t}_2, \ldots$, etc. Now to the definitions of truth, falsity, and negation in a given interpretation $\mathcal{I}$ with domain D:

Definition 1. $P(\underline{t}_1,\ldots,\underline{t}_n)$ is true in $\mathcal{I}$ iff $\langle \mathbf{t}_1,\ldots, \mathbf{t}_n \rangle \in R_1$; in symbols: $\models_{\mathcal{I}} P(\underline{t}_1,\ldots,\underline{t}_n)$.

Definition 2. $P(\underline{t}_1,\ldots,\underline{t}_n)$ is false in $\mathcal{I}$ iff $\langle \mathbf{t}_1,\ldots, \mathbf{t}_n \rangle \in R_2$; in symbols $\models_{\mathcal{I}} \neg P(\underline{t}_1,\ldots,\underline{t}_n)$.

Definition 3. $\neg P(\underline{t}_1,...,\underline{t}_n)$ is true in $\mathcal{I}$ iff $\langle \mathbf{t}_1,..., \mathbf{t}_n \rangle \in R_3$; in symbols $\models_{\mathcal{I}} \neg P(\underline{t}_1,...,\underline{t}_n)$.

Definition 4. $\neg P(\underline{t}_1,...,\underline{t}_n)$ is false in $\mathcal{I}$ iff $\langle \mathbf{t}_1,..., \mathbf{t}_n \rangle \in R_4$; in symbols $\in_{\mathcal{I}} \neg P(\underline{t}_1,...,\underline{t}_n)$.

The set-theoretic relations R_i are not fixed beforehand; they can be pairwise disjoint, intersect, be included one in another, etc. Thus we can have the following cases for a given predicate P.

1. $R_1 \cup R_2 = D^n$; $P(\underline{t}_1,...,\underline{t}_n)$ is either true or false.

2. $R_1 \cup R_2 \subset D^n$; for some n-tuples $\langle t_1, ..., t_n \rangle$, $P(\underline{t}_1,...,\underline{t}_n)$ is neither true nor false.

3. $R_1 \cup R_3 = D^n$; $P(\underline{t}_1,...,\underline{t}_n)$ is true or $\neg P(\underline{t}_1,...,\underline{t}_n)$ is true.

4. $R_1 \cup R_2 = D^n \wedge R_1 \cap R_2 \neq \varnothing$; $P(\underline{t}_1,...,\underline{t}_n)$ is true, false, or antinomic.

5. $R_2 \subseteq R_3$; if $P(\underline{t}_1,...,\underline{t}_n)$ is false, $\neg P(\underline{t}_1,...,\underline{t}_n)$ is true, the converse is not necessarily the case.

6. $R_3 \subseteq R_2$; if $\neg P(\underline{t}_1,...,\underline{t}_n)$ is true, $P(\underline{t}_1,...,\underline{t}_n)$ is false, the converse is not necessarily the case.

7. $R_1 \cap R_3 \neq \varnothing$; for some n-tuples $\langle \underline{t}_1,...,\underline{t}_n \rangle$, $P(\underline{t}_1,...,\underline{t}_n)$ is true and so is $\neg P(\underline{t}_1,...,\underline{t}_n)$.

8. $R_2 \cap R_4 \neq \varnothing$; for some n-tuples $\langle \underline{t}_1,...,\underline{t}_n \rangle$, $P(\underline{t}_1,...,\underline{t}_n)$ is false and so is $\neg P(\underline{t}_1,...,\underline{t}_n)$.

9. $R_2 = R_3$; $P(\underline{t}_1,...,\underline{t}_n)$ is false iff $\neg P(\underline{t}_1,...,\underline{t}_n)$ is true.

10. $R_3 \cup R_4 \subset D^n$; for some n-tuples $\langle \underline{t}_1,...,\underline{t}_n \rangle$, $\neg P(\underline{t}_1,...,\underline{t}_n)$ is neither true nor false.

As these examples show, there is no truth and falsity in the abstract but only in reference to a specific interpretation: R_1 to R_4 and their set-theoretic relationships can be assigned to P very differently in different domains.

Having considered the atomic predicate formulas, we can now use Definitions 1 and 2 to extend the notion of satisfiability to all positive well-formed formulas.

Definition 5. $A(x_1,..., x_n)$ is a well-formed positive predicate formula iff it is formed in accordance with the usual rules of formation and neither "$\neg$" nor the existential quantifier "$\exists x_i$" occur in the formula.

Definition 6. A well-formed positive predicate formula $A(x_1,...,x_n)$ is satisfiable in a given interpretation $\mathcal{I}$ iff for some n-tuple $<t_1,...,t_n>$, $A(x_1,...,x_n)$ meets the usual definition of satisfiability. Then, $A(x_1,...,x_n)$ is true in $\mathcal{I}$ iff it is satisfied by all n-tuples in $\mathcal{I}$, and logically valid iff it is true in all interpretations.

As with the truth of negation in the propositional calculus, expressions involving negation and the existential quantifier are to be considered satisfiable, true, or logically valid in an ad hoc manner. Although much of the meaning of classical existential quantification is meant to be retained, the usual definition "$\exists xA(x)$ stands for $\neg\forall x\,\neg A(x)$" does not hold in this work, that is, $\exists x$ is to be taken as a primitive operator. The usual relation between universal and existential quantification "$\forall xA(x) \Rightarrow \exists xA(x)$," which is intuitionistically acceptable, will be the case here occasionally but not always. It is possible to have $\models\forall xA(x)$ and $\models\neg\exists xA(x)$. Since intuitionistically $\forall xA(x)$ must be constructively determined, it stands to reason that $\exists xA(x)$ follows, i.e., that what is true for all must be true for some. But if nonconstructive methods are accepted (excluded middle, axiom of choice, and the like), $\forall xA(x)$ may be deducible without our having any method to find an x such that $A(x)$, thus opening the possibility that no such x exists.

Here, then, asserted formulas involving existential quantification will each have the status of a proper axiom. Only the positive fragment of the antinomic predicate calculus will retain its classical deductive generality, it being understood that "true" in Definition 6 above includes the case in which a positive well-formed formula is both true and false.

Further, since we do not have the classical satisfiability rules for existential quantification, it is possible to have $\models\exists xA(x)$ and $\models\neg\exists xA(x)$: an x that satisfies $A(x)$ may exist and not exist. For example, to say that a function f mapping the set A onto the set B in a one-to-one manner exists means: for every $a\in A$ there is a unique $b\in B$ such that $(a,b)\in f$. Yet, if there is an $a\in A$ such that not only $(a,b)\in f$ but also $\neg(a,b)\in f$, then we must conclude that f simultaneously exists and does not exist, and that the image $f(a)$ of a exists and does not exist at the same time. The existence or not of f means, precisely, the membership or not of the appropriate ordered pairs (a,b) to f. In general, $\forall xA(x)$ may be true in an abstract sense without having any concrete individual x satisfying $A(x)$; in these cases, $\neg\exists xA(x)$ is asserted — a possibility that is not counterintuitive but rather is the natural result of the acceptance of nonconstructive methods.

Definition 7. The model for a well-formed predicate formula A is any interpretation $\mathcal{I}$ in which A is true or antinomic according to the following. (i) If A is positive, A is true in an interpretation $\mathcal{I}$ iff A fulfills in $\mathcal{I}$ the usual definition of truth restricted to such positive formulas; note that A may also be false , i.e., antinomic. (ii) If A is negative then A is asserted as true, antinomic, or logically valid by specific designation.[10] Note that for both the positive and negative formulas we can have A (i) antinomic for some valuations in the given interpretation $\mathcal{I}$; (ii) antinomic for all valuations in $\mathcal{I}$; in other words, true and false in $\mathcal{I}$ or fully antinomic in $\mathcal{I}$; (iii) false for some valuations in $\mathcal{I}$; (iv) false for all valuations in $\mathcal{I}$; (v) logically false, i.e., false in all interpretations, a notion that is independent of negation since, again, false is not necessarily "not true"; and (vi) logically antinomic, i.e., fully antinomic (true and false) in all interpretations.

As for syntax, the axioms for the positive fragment of the predicate calculus are the same classical positive axiom schemes:

(i) $\forall x A(x) \Rightarrow A(t)$, with t a term free for x in $A(x)$.

(ii) $\forall x(A \Rightarrow B) \Rightarrow (A \Rightarrow \forall x B)$, with A having no free occurrence of x.

No axiom scheme for negative formulas will be added: negative formulas will be asserted as needed, not inferred, much as one chooses proper axioms for a given first-order theory.

In addition, let us postulate the rule of inference of generalization, from A, $\forall x A$ follows; in symbols, $A \vdash \forall x A$, where A is a positive well-formed formula (only positive formulas can be inferred).

An alternative way to define a positive predicate logic would be to retain $\wedge$, $\vee$, $\Rightarrow$, $\Leftrightarrow$, but substitute the universal quantifier with the existential one. Whereas the positive logic with $\forall x$ exclusively is a logic of generalities, the positive logic with $\exists x$ exclusively is a logic of particular cases. In the latter, the axiom schemes would be different, including, for example, $A(x) \Rightarrow \exists x A(x)$; also, the rule of generalization would be replaced by the introduction of the existential quantifier as follows: if B does not contain x free, then $A(x) \Rightarrow B \vdash \exists x A(x) \Rightarrow B$. The positive predicate logic thus obtained would be different from the previous one, of course, and the negative formulas would be those in which "$\neg$" or "$\forall x$" occur. Once more, the negative fragment of this predicate calculus would share some of the characteristics of a first-order

theory, with any asserted negative formula having the status of an ad-hoc axiom. These axioms would be all the negative theorems since, again, we would have no rule of inference for negative formulas, a situation that is similar to the way in which one defines a complete theory by postulating as axioms all the well-formed formulas true in a given model.

Whether one selects $\forall x$ or $\exists x$ as the positive quantifier, neither one can be defined in terms of the other and negation in the usual way. We have already made the point that, with $\forall x$ as the positive quantifier, we cannot automatically transfer the validity of a property for a whole class of individuals to the validity of that property for a single specific individual. In the second case, with $\exists x$ as the positive quantifier, it is possible to have $\models \exists x A(x)$ and $\models \forall x \neg A(x)$, that is, local validity does not necessarily have any of the usual consequences for global validity. Here, we shall stay with the first case, i.e., with $\forall x$ as positive, and add the completeness theorem of the predicate calculus as a meta-axiom for all well-formed formulas, positive and negative. Thus A is logically valid if and only if it is a theorem, $\models A$ iff $\vdash A$. As a result, a negative formula that is true or antinomic in all interpretations is automatically an axiom of the predicate calculus. For the negative fragment of the predicate calculus, then, semantics fully determines the syntax; the positive fragment remains close to the classical two-way form of completeness (allowing, of course, for the possibility of true formulas that are also false).

Let us pause now to elaborate on the meaning of the existential quantifier in the context of this antinomic logic. Informally, we shall characterize the existential quantification $\exists x P(x)$ not as the disjunction $P(x_1) \vee P(x_2) \vee \ldots \vee P(x_k)$ (extendable to an infinite number of disjuncts), nor as the class $\{x_i : P(x_i)\}$ of all individuals x_i for which $P(x_i)$ holds, but as one single individual choice from the collection of all individuals satisfying $P(x)$: in symbols, $\iota x P(x)$, extending the meaning of the iota symbol (introduced by Russell for the description of individuals) from referring only to the unique x such that $P(x)$[11] to referring to a nonspecified individual chosen from $\{x_i : P(x_i)\}$. Having put aside the usual definition "$\exists P(x)$ stands for $\neg\forall x\neg P(x)$" allows us to map the well-formed formula $\exists x P(x)$ into one single individual as the formula's meaning (if no x satisfies P(x) in a given interpretation, $\iota x P(x)$ is the empty set). All this is similar to the above mentioned informal characterization of set given by Cantor; i.e., it is intended to provide an intuitive justification for the cleavage we have drawn between $\forall x$ and $\exists x$. Note that since $\models \forall x P(x)$ and not-$\models \exists x P(x)$ are simultaneously possible, a property can be generally true without being true specifically: $\models \forall x P(x)$ is

compatible with not being able to find an individual value a for x such that $\models P(a)$. Let us look at an example of this situation, still informally.

Let $C(x)$ be a function that determines the cardinality of a set x, that is, a set $|x|$ that can be defined with or without the axiom of choice. Let $x \equiv y$ indicate that x can be mapped in a one-to-one manner onto y. We shall assume that there may be several such cardinality functions, but that if C and C' are any two of such functions, then $C(x) \equiv C'(x)$. Assume that universal and existential quantification is restricted to these cardinality functions. Then $\models \forall C((C(x)=C(y) \Leftrightarrow C(x) \equiv C(y))$ obtains, but not-$\models \exists C((C(x)=C(y) \Leftrightarrow C(x) \equiv C(y))$ can be the case at the same time, since there are models of Zermelo-Fraenkel's set theory with a proper class of atoms in which no function C can be defined for all x with the property $C(x)=C(y) \Leftrightarrow C(x) \equiv C(y)$.[12]

It is advisable now to point out explicitly some classical theorems and metatheorems which will hold in some cases but definitely not in all. For example, classically, if in any theory T it is the case that $\vdash A \vee B$ implies $\vdash A$ or $\vdash B$, then and only then T is syntactically complete, i.e., $\vdash A$ or $\vdash \neg A$ for any well-formed formula A. The proof of this equivalence requires excluded middle, contraposition, and the tautology $\neg A \Rightarrow ((A \vee B) \Rightarrow B)$, all of which are not valid here, both in the object language and in the metalanguage. In addition, the following negative formulas can be true, or false, or both, or neither: $\neg(A \wedge \neg A \Rightarrow B)$, $\neg((A \Rightarrow B) \Rightarrow ((A \Rightarrow \neg B) \Rightarrow \neg A))$, $\neg(A \vee \neg A)$, $\neg(\neg \neg A \Rightarrow A)$, $\neg(A \Rightarrow \neg \neg A)$, $\neg((\neg A \Rightarrow \neg B) \Rightarrow (B \Rightarrow A))$. There will be cases in which $A \vee \neg A$ is a good choice for some A's and $\neg(B \vee \neg B)$ is a good choice for some B's. The same applies to the other formulas just listed. In particular, the law of excluded middle, a negative metatheorem not itself responsible for contradictions and not assumed here in general, as we mentioned, could be assumed in particular to make room for the conclusion that every real number has a decimal expansion, even though Brouwer actually exhibited a definite number for which it is not known if there is a first digit in its decimal expansion.[13] The prime ideal theorem, used in the proof of Gödel's completeness theorem, is also a negative metatheorem which will not be assumed here, although the completeness theorem will be assumed in general as a meta-axiom for every first-order theory.

Finally, as already stated, both $\exists x A(x)$ and $\neg \exists x A(x)$ may be true, and hence axioms. But it is also possible that not-$\models \exists x A(x)$ and not-$\models \neg \exists x A(x)$ are the case, together with not-$\vdash \exists x A(x)$ and not-$\vdash \neg \exists x A(x)$; that is, $\exists x A(x)$ and its negation are neither true nor false. Thus, instead of saying that the sentence "there is a white unicorn" is false because unicorns

do not exist in reality, here, precisely because unicorns cannot be found in reality and therefore a white one cannot be selected, the sentence is neither true nor false. If A. Robinson's definition of the complete diagram of a given model is extended to include not only those sentences which are true in that model but also those which are antinomic in the model, then we must also exclude from the diagram not only the simply false sentences but also those which are neither true nor false in the model. Note that whereas the positive fragment of this complete diagram can be considered deductively predetermined, the negative fragment is always open to enlargement when negative formulas are axiomatically added as needed (again, the interpretation of negative atomic formulas does not predetermine the truth or falsity of the compound ones).

§7. Equality as an antinomic predicate.

The motivation behind antinomic logic lies in the conviction that, irreducibly, there is identity in difference in many realms, including nature. As a consequence, $\neq x=y$ and $\exists x=y$ together must be considered possible for some values of x and y. Since here equality is to be defined in terms of membership, we shall not add equality as a primitive antinomic predicate because it will turn out to be antinomic as a derived one.[14]

§8. Other kinds of antinomicity.

It is a mistake to think that antinomicity is exclusively caused by negation: negationless systems can harbor their own forms of antinomicity. Nor must antinomic statements be defined in terms of truth and falsity. Any kind of opposition can produce its own form of antinomicity — whole and part, one and many, and a host of other contrasting concepts which do not necessarily involve negation and which can be considered independently of truth and falsity. Here we shall restrict ourselves solely to antinomic sentences and formulas in the sense in which they have been introduced above.

III. Antinomic Set Theories

§9. Antinomic membership.

Some sets will be antinomic in the sense that they belong and do not belong to another set, that is, $\neq x\in y \wedge x\notin y$ and $\exists x\in y$, regardless of whether $\exists x\notin y$ or not-$\exists x\notin y$, abbreviated $x\in\notin y$,

which will be read "x is an antinomic member of y," or "y contains x as an antinomic member." The set y need not be an antinomic member of another set. In what follows, some sets will be antinomic members of other sets and nonantinomic members of still others; some sets will not be antinomic members of any other set; and other sets will be antinomic members of any set to which they belong. Symmetrically, some sets may have some antinomic members and some nonantinomic ones; others may not have a single antinomic member; and still others may have only antinomic members.

The language of set theory will include variables x, y, z, u, v, w, x_1, x_2, x_3, ..., to range in given domains, and also constants a, b, c, a_1, a_2, a_3, ..., to represent single fixed sets. We shall postulate set-theoretic completeness below; as a consequence, if $a \in b$ is true ($\models a \in b$), then $a \in b$ is an axiom or a theorem ($\vdash a \in b$), and vice versa. Also, if $\models a \notin b$, then $\vdash a \notin b$, and vice versa.

In addition to the notation $a \in\!\!\not\in b$ already introduced, we shall represent by $a \in\!\!\!\in b$ the case in which $\models a \in b$ but not-$\models a \notin b$ and not-$\vDash a \in b$, regardless of whether $\vDash a \notin b$ or not-$\vDash a \notin b$. The metamathematical negation "not-$\models A$" stands for "A is not true," and is equivalent metamathematically to not-$\vdash A$, "A is not provable," given completeness; not-$\vDash A$ means A is not false. Therefore, $a \in\!\!\!\in b$ implies that the sentence $a \notin b$ is neither true nor a theorem. Finally, let us use $a \notin\!\!\!\in b$ to represent the case in which $\models a \notin b$ but not-$\models a \in b$ and not-$\vDash a \notin b$, regardless of whether $\vDash a \in b$ or not-$\vDash a \in b$ ($a \in b$ is neither true nor is it therefore a theorem). As determined in Part II, the metamathematical negation "not" must be distinguished from the formal negation "$\neg$" in that the metamathematics of antinomic set theory is not antinomic in the following sense: although A may be true and false, it is not the case that A is and is not true ($\models A$ and not-$\models A$); or, correspondingly, that A is both provable and unprovable ($\vdash A$ and not-$\vdash A$).

Given an arbitrary set b and a member a both in a given universe w such that $a \in w$ and $b \in w$ or b is included in w (see Definition 10 below), we shall assume that it is always determined which of these three mutually exclusive cases is in order: (i) $a \in b$, (ii) $a \notin\!\!\!\in b$, or (iii) $a \in\!\!\not\in b$. These cases are relative to the given universe w; that is, $a \in b$ in w_1 is compatible with $a \notin\!\!\!\in b$ in w_2, and with $a \in\!\!\not\in b$ in w_3: although the antinomicity of membership is a matter between a set a and the set b to which a belongs, it is dependent on the universe w in which both are being considered. Further, within the same universe w, a may be an antinomic member of b and a nonantinomic member of a proper subset or a proper superset of b. In a

relative universe w, antinomicity is strictly an internal relation between a and b, a complex kind of membership and not a property that is intrinsic to the member a or the set b. Thus, we can say that a is a "circumstantially" antinomic member of b which can be "de-antinomized" by changing the universe w, or simply by considering a as a nonantinomic member of another set c in the same universe. Antinomicity is a variable, not an absolute condition.

§10. Axioms for an antinomic set theory AS_1 based on membership.

The presence of antinomic sets in a given universe w forces us to review the usual axioms to make room for the new cases. Let us begin by considering equality.

Definition 8. $y=z$ stands for $\forall x(x\in y\Leftrightarrow x\in z)$. If y and z are included in w (see Definition 10 below), this definition thoroughly defines equality in w. If $y\in w\wedge z\in w$, the following becomes necessary.

Axiom 1. $y=z\Rightarrow\forall u(y\in u\Leftrightarrow z\in u)$. This extensionality in terms of $\in$ obtains in the relative universe w within which x, y, z, and u are considered. But extensionality determines uniqueness of sets only insofar as the all-inclusive membership $\in$ is concerned — that is, uniqueness must be understood as "modulo" antinomicity, disregarding the branching of $x\in y$ into either $x\in y$ or $x\in\not\in y$.

Each specific unique set in this sense will be represented by a constant a, b, c, etc., as mentioned. But although $y=z$ is an equivalent relation that implies that y and z have the same $\in$-members in a given universe w and may be represented therefore by the same constant a, because the type of membership of x to z may vary from universe w_1 to universe w_2, then if $y=z$ in both w_1 and w_2, the following two cases are compatible with $x\in a$ (the meaning of $z\subseteq w$ is given in the usual way in Definition 10 below).

$\vDash x\in w_1\wedge(a\in w_1\vee a\subseteq w_1)\wedge x\in a$, and $\vDash x\in w_2\wedge(a\in w_2\vee a\subseteq w_2)\wedge x\in\not\in a$.

Definition 9. $y\neq z$ stands for $\exists x((x\in y\wedge x\in\not z)\vee(x\in\not y\wedge x\in z))\vee\exists u((y\in u\wedge z\in\not u)\vee(y\in\not u\wedge z\in u))$.

Let us then distinguish the following particular cases: (i) $y\mathbf{=}z$ stands for $y=z\wedge\forall x(x\in y\Leftrightarrow x\in z)\wedge\forall u(y\in u\Leftrightarrow z\in u)\wedge\forall x(x\in\not y\Leftrightarrow x\in\not z)\wedge\forall u(y\in\not u\Leftrightarrow z\in\not u)$.

(ii) $y_a=z$ stands for $y=z\wedge\exists x(x\in\not y)\wedge\forall v(v\in z\Rightarrow v\in z)$. Symmetrically, the meaning of $y=_a z$ is obvious.

(iii) $y_a=_a z$ stands for $y=z\wedge\exists x(x\in\not\in y)\wedge\exists v(v\in\not\in z)\wedge x\neq v$.

(iv) $y^a=z$ stands for $y=z\wedge\exists u(y\in\not\in u)\wedge\forall v(z\in v\Rightarrow z\in v)$. The meaning of $y=^a z$ is obvious.

(v) $y^a=^a z$ stands for $y=z\wedge\exists u(y\in\not\in u)\wedge\exists v(z\in\not\in v)\wedge u\neq v$.

(vi) $y^a{}_a=^a{}_a z$ stands for $y_a=_a z\wedge y^a=^a z$.

These different cases show that equality is a type of equivalence relation that can be interpreted as strict regular identity in terms of $\in$ and $\in\not$ if and only if $y=z$. Thus, even if $y=z$ obtains in all universes w, the kind of extensionality of y and z may vary from one universe to another, say, $y_a=z$ in w_1 and $y=^a z$ in w_2, even if y and z are not only equal but have exactly the same $\in$-members in w_1 and w_2. Also, since $y\neq z$ obtains if $\exists x(x\in y\wedge x\in\not z)$, then $y=_a z$ entails $y=z\wedge y\neq z$, i.e., equality is antinomic in such cases. In particular, two relative universes w_1 and w_2 may be equal and different at the same time if, say $w_1=w_2$ but $w_{1a}=w_2$ specifically. All this necessarily affects the application of any of the forthcoming axioms in which the existence of a set is relativized to a given universe.

§11. Inclusion.

Let us now define inclusion in the usual way.

Definition 10. $y\subseteq z$ stands for $\forall x(x\in y\Rightarrow x\in z)$, with proper inclusion, $y\subset z$, meaning $y\subseteq z\wedge\exists x(x\in z\wedge x\in\not y)$.

With this definition, $y=z$ is compatible with $y\subset z$ and $z\subset y$; obviously, $y_a=z$ implies $y\subset z$. We shall distinguish the following cases:

(i) $y_a\subset z$ for $y\subset z\wedge\exists x(x\in\not y\wedge x\in z)$.

(ii) $y\subset_a z$ for $y\subset z\wedge\exists x(x\in y\wedge x\in\not y)$.

(iii) $y_a\subset_a z$ for $y_a\subset z\wedge y\subset_a z$.

(iv) $y\subseteq z$ for $y\subseteq z\wedge\forall x(x\in y\Rightarrow x\in y\wedge x\in z)$.

To repeat, note that membership of a set x to a set y, being strictly a matter between x and y relative to the universe in which they are considered, has nothing to do with the kind of membership of x to the proper subsets of y or to the proper supersets of y. That is, if $x\in u\subset y\subset z$, it is possible to have $x\in\not u\wedge x\in y\wedge x\in\not z$, etc. In addition, if we change the relative

universe w in which x and y are considered, $x \in y$ may become $x \in \notin y$. As a particular case, if in any universe w a set x is an antinomic member of every set to which it belongs, $\forall w \forall y(x \in w \wedge (y \in w \vee y \subseteq w \Rightarrow (x \in y \Rightarrow x \in \notin y)))$, we can represent this situation with the one-place predicate Ant(x), defined by the last formula which reads "x is universally antinomic."

Finally, antinomicity makes possible mutual proper inclusion. In other words, if proper inclusion is taken as the set-theoretic meaning of the phrase "being a part of," then it is possible for two sets to each be a part of the other. Further, we can even say that the whole can be part of the part, i.e., $y \subset z \subset y$ if $y = z$ and $\exists x(x \notin y \wedge x \in z) \wedge \exists v(v \in y \wedge v \notin z)$. This is also the case if we use the expression "being a part of" in the set-theoretic sense of being a member of, that is, $x \in y \in x$. We shall not assume the axiom of foundations that rules out $x \in x$, $x \in y \in x$, etc., hence, $x \notin x$, $x \in x$, $x \in \notin x$, $x \in y \in x$, etc., all are distinctly possible.

§12. Axiom of comprehension.

One good mathematical reason for building antinomic set theories is to retrieve Cantor's comprehension axiom in its original unrestricted form; this return to "Cantor's paradise" would have significant consequences for the mathematical usefulness of such theories. Here, however, since we want to relativize membership as much as possible, we shall use an antinomic version of Zermelo's axiom of separation, the standard form of which is expressible as follows: Given a set y and an arbitrary set-theoretic formula $A(x)$ in which y does not occur and x is a free variable, there exists a set z such that $x \in z \Leftrightarrow (x \in y \wedge A(x))$. In this form, several possibilities are in order in accordance with the two mutually exclusive meanings of membership, that is, whether $x \in y$ is interpreted as $x \in y$ or $x \in \notin y$, and whether $x \in z$ is interpreted as $x \in z$ or $x \in \notin z$. To leave the ambiguity unresolved would mean that z would not be strictly unique; in effect, we could have as many z's as there are ways in which these four possibilities can be combined. In order to make z uniquely determined in each relative universe w, we postulate specifically:

Axiom 2: $\forall w \forall y(y \in w \vee y \subseteq w \Rightarrow \exists z(z \subseteq y \wedge \forall x(x \in w \Rightarrow ((x \in z \Leftrightarrow A(x) \wedge x \in y) \wedge (x \in \notin z \Leftrightarrow A(x) \wedge x \in \notin y)))))$, in which $A(x)$ does not involve any of the quantified variables w, y, and z, and in which x is a free variable. In other words, the kind of membership of x to z is determined by the kind of membership of x to y. The notation $z = \{x : (x \in w \wedge x \in y) \wedge (y \in w \vee y \subseteq w) \wedge A(x)\}$ is now in order, and its meaning is unambiguously determined by Axiom Scheme 2. If w is fixed exclusively, then the expression

$z=\{x{:}x\in y\wedge A(x)\}$ suffices; and if in addition y is w, then $z=\{x{:}A(x)\}$ suffices, and z will gather those sets x which are members of w and satisfy $A(x)$, with w fixed.

§13. Russell's paradox.

If $A(x)$ is $x\notin x$, then $z=\{x{:}x\in y\wedge x\notin x\}$. If $z\in y$ and $z\notin z$, then $z\in z$, that is, $z\in\notin z$ — z is an antinomic member of itself. If y is also an antinomic member of itself, then $y\in z$, although $y\notin z$ remains undetermined. If $\forall x(x\in y\Rightarrow x\notin x)$, then $z=y$, even if $x\in\notin x$ for some x. If, on the other hand, there is an x such that $x\in x$, then $z\subset y$. In any event, Russell's paradox is harmless even if it leads to contradictions.

§14. Other axioms and the Boolean operations.

The following axioms are not all independent and each is relativized to a circumstantial universe w in which the sets involved are (i) members of w, (ii) members of members of w, or (iii) subsets of w. We shall not make this relativization to w explicit in all the axioms nor for all the sets, and will assume w fixed when it does not occur in the expressions that follow. Note once more that the kind of membership of x to w does not determine the kind of membership of x to any member of w.

Axiom 3. $\forall y\forall z(y\in w\wedge z\in w\Rightarrow \exists u(u\in w\wedge\forall x(x\in u\Leftrightarrow x=y\vee x=z)))$. Pairing.

Axiom 4. $\forall y(y\in w\Rightarrow \exists u(u\in w\wedge\forall x(x\in u\Leftrightarrow x\subseteq y)))$. Power set.

Axiom 5. $\forall y\exists u\forall x(x\in u\Leftrightarrow\exists z(x\in z\wedge z\in y))$. Union.

Axiom 6. $\exists y(\forall x(x\notin y))\wedge\forall y\forall z(\forall u(u\notin y)\wedge\forall v(v\notin z)\Rightarrow y=z)$. Null set.

Definition 11. $\{y,z\}$ represents the unique set modulo antinomicity determined by Axiom 3; $\{y\}$ stands for $\{y,y\}$. $\mathscr{P}(y)$ represents the unique power set modulo antinomicity determined by Axiom 4. The expression "modulo antinomicity" already used in connection with equality here means, precisely, that in applying Axioms 3, 4, and 5 as well, two sets u and u' may exist in each of these three cases that satisfy the axiom but such that $x\in u\wedge x\in\notin u'$, say, and yet, $u=u'$ in each case. Finally, $\varnothing$ represents the unique null set; $\varnothing$ does not have antinomic members, although it may be the antinomic member of other sets; further, $\varnothing\in a$ may be true in the universe w_1 but $\varnothing\in\notin a$ may also be true in w_2.

Note that the various kinds of inclusion, together with Axiom 2, allow us to distinguish special power sets as follows.

(i) $\mathscr{P}(y)=\{x:x\in\mathscr{P}(y)\wedge x\subseteq y\}$,

(ii) $\mathscr{P}_a y=\{x:x\in\mathscr{P}(y)\wedge x\subseteq_a y\}$,

(iii) ${}_a\mathscr{P}(y)=\{x:x\in\mathscr{P}(y)\wedge x_a\subseteq y\}$,

(iv) ${}_a\mathscr{P}_a y=\{x:x\in\mathscr{P}(y)\wedge x_a\subseteq_a y\}$.

The kind of membership of x to $\mathscr{P}_a y$, etc., is determined by the kind of membership of x to $\mathscr{P}(y)$ in accordance with Axiom 2.

Axiom 2 also guarantees the existence of the usual set-theoretic operations, but some restrictions should apply on the possible kinds of membership. For the case of intersection, for example, the usual Boolean definition $x\in y\cap z\Leftrightarrow x\in y\wedge x\in z$ will hold in general, but whether $x\in y\cap z$ or $x\in\not\in y\cap z$ will depend on the kind of membership of x to the universe w in which the intersection is considered. To make certain that the operations are single-valued in each universe, we then define the following:

Definition 12.

(i) $y\cap_w z=\{x:x\in y\wedge x\in z\}$, which implicitly means $((x\in w\Rightarrow x\in y\cap z)\wedge(x\in\not\in w\Rightarrow x\in\not\in y\cap z))$. The kind of membership of x to y and to z is irrelevant; note also that y and z are each either a member or a subset of w, given that Axiom 2 relativizes comprehension to a fixed universe w. The subindex w in $y\cap_w z$ can be dropped when w is taken for granted. In fact, given the final remark in §12, $y\cap z=\{x:x\in y\wedge x\in z\}$ is sufficient as a definition of intersection if we take $A(x)$ to mean $x\in y\wedge x\in z$ with y and z as fixed parameters.

(ii) $y\cup_w z=\{x:x\in y\vee x\in z\}$, which implicitly means $((x\in w\Rightarrow x\in y\cup z)\wedge(x\in\not\in w\Rightarrow x\in\not\in y\cup z))$.

(iii) $y'_w=\{x:x\in\not\in y\}$, which implicitly means $((x\in w\Rightarrow x\in y')\wedge(x\in\not\in w\Rightarrow x\in\not\in y'))$. Again, note that the kind of membership of x to the complement of y is determined not by the kind of nonmembership of x to y but by the kind of membership of x to w. Thus, the two mutually exclusive cases follow: first, if $x\in w\wedge x\in y'$, then $x\in y'$, whether $x\in\not\in y$ or $x\in\not\in y$; second, if $x\in\not\in w\wedge x\in y'$, then $x\in\not\in y'$, whether $x\in\not\in y$ or $x\in\not\in y$. The expression $y'_w=\{x:x\in\not\in y\}$ implicitly assumes this distinction.

(iv) $S_w y=\{x: x\in y \vee x=y\}$, where $x\in S_w y$ if $x\in w$, and $x\in\not\in S_w y$ if $x\in\not\in w$. In addition, $\mathbf{S}_w y=\{x: x\in y \vee x=y\}$ where $y\in w$ and hence $y\in \mathbf{S}y$ also. If w is fixed, we simply write Sy and $\mathbf{S}y$. For Sy, and $\mathbf{S}y$ in particular, we shall assume $\models(y=z\Rightarrow Sy=Sz)\wedge(\mathbf{S}y=\mathbf{S}z\Rightarrow y=z)$.

(v) Nat(x) iff $x=\varnothing\vee(x=\mathbf{S}y\wedge\text{Nat}(y))$, x is a natural number.

Axiom 7. $\exists y(\varnothing\in y\wedge\forall x(x\in y\Rightarrow \mathbf{S}x\in y))$. Existence of an infinite set with an infinity of nonantinomic members. The axiom also guarantees the existence of an infinity of natural numbers.

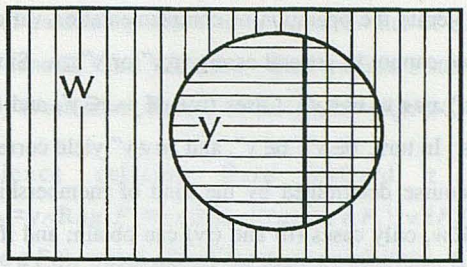
Axiom 8 (meta-axiom). The antinomic set theory AS_1 satisfies completeness in the sense that A is an axiom or a theorem of AS_1 if and only if it is true in all models of AS_1: $\vdash A$ iff $\models A$. It should be re-emphasized that $\models A$ includes these two mutually exclusive cases: (i) A is simply true, $\models A$ but not-$\boxminus A$, and (ii) A is true-and-false, $\models A$ and $\boxminus A$, noting that A could be simply true in one model and antinomic in another despite being true in all models of AS_1. For positive formulas in AS_1 the only change with respect to the classical situation is the addition of semantic antinomicity in some cases. For negative formulas in AS_1 the application of Axiom 8 is ad hoc and goes from semantics to syntax. Again, the positive diagram of a given model of AS_1 is predetermined by the axioms. The negative diagram, i.e., the collection of all negative formulas true or antinomic in such a model, remains incomplete and open to successive additions.

Axiom 8 does not imply that AS_1 is syntactically complete, although the existence of a complete extension of AS_1 can certainly be assumed. Since AS_1 is far from having a recursive set of axioms, Gödel's first incompleteness theorem does not apply; but even if AS_1 could be presented as an axiomatizable extension of formal number theory, once one gives up the premise of consistency Gödel's second incompleteness theorem does not apply either.

§15. Relative complementation and Venn diagrams.

Definition 13. $z\text{-}y=\{x: x\in z\wedge x\not\in y\}$, complement of y relative to z for all sets y and z that are either members or subsets of the implicit universe w.

Because of antinomicity, some members x of a relative universe w, which in turn contains y as a subset, may belong to y and to its complement. The Venn diagram for the complement of y (y represented by the horizontally shaded area inside the circle) would look like the following vertically shaded area.



That is, y and y' intersect, and the members of this nonempty intersection are those within the doubly shaded area inside the circle. The area inside the circle not in $y \cap y'$ corresponds to $y\text{-}y' = \{x : x \in y \wedge x \notin y'\} = \{x : x \in y\}$.

Note that $x \in \not\in y$ is compatible with $x \notin y'$ (if $x \in w$), and precisely because the antinomic member of a set is not necessarily the antinomic member of its complement, $y \neq y''$ is possible. In effect, y'' may be a proper subset of y if y' has no antinomic members, but if it does, then again y' and y'' would intersect and y'' would not be contained in y; if $x \in y'$, then $x \notin y''$. Also, $z \subset y$ does not imply $y' \subset z'$, since for the same x we may have $x \in \not\in y$ and $x \in z$, i.e., $x \in y' \wedge x \notin z'$. Further, since $x \in \not\in y \wedge x \in \not\in z \wedge x \in y \cap z$ is possible (if $x \in w$), then $x \notin (y \cap z)'$ even though $x \in y' \wedge x \in z'$. If we define in the usual way the generalized intersection $\cap_i y_i$ (relative to a universe w) of a family of sets (each included in w) indexed by an index set I, then if there is a set x such that $\forall i (i \in I \Rightarrow x \in \not\in y_i)$, $x \in \cap_i y_i$ but also $x \in \cap_i y_i'$. If the relative universe w contains a single antinomic member $x \in \not\in w$, then the complement of w is not empty. If $y \subseteq w$ and all the members of y are antinomic, $y \subseteq y'$; also, many subsets y_i of w could have the same complement, and if for each y_i all its members were antinomic, then $y_i' = w$ for all i. In extreme cases, if $y \subseteq w$ and $\forall x (x \in y' \Rightarrow x \in y)$, then $y' \subseteq y = w$, and if $\forall x (x \in y' \Leftrightarrow x \in y)$, then $y = y' = w$.

Because the laws of double negation are not valid, the logical De Morgan laws do not obtain, and neither do the set-theoretic De Morgan laws. For example, (i) $(y \cap z)' \subset y' \cup z'$ and (ii) $(y \cap z)' \supset y' \cup z'$ are both possible cases. To see this, consider that the kind of membership of a set x to y', z', $(y \cap z)'$, and $(y \cup z)'$ is determined according to Axiom 2 and Definition 12 by the kind of membership of x to the relative universe w of which y and z are members or subsets. The proper inclusion (i) is possible because members x of $y' \cup z'$ may not be members of $(y \cap z)'$ if the following obtains: if $x \in w$, $x \in \not\in y$, $x \in z$, then $x \in y'$ and $x \in y' \cup z'$, but $x \in y \cap z$, hence $x \notin (y \cap z)'$. As for (ii), it may obtain if $x \in \not\in w$, $x \in y$, and $x \in z$, for then $x \in \not\in y \cap z$ but $x \notin y'$ and $x \notin z'$, hence $x \notin y' \cup z'$. Also, neither $(y \cap z)'$ nor $y' \cup z'$ may be included in the other.

As hinted, if we iterate the operation of complementation various cases are possible. (i) If $x \notin y$ and $x \in \notin y'$, we cannot in general assert $y \subseteq y''$ or $y'' \subseteq y$. Similar situations arise if (ii) $x \in \notin y \wedge x \in y'$, and (iii) $x \in \notin y \wedge x \in \notin y'$. Cases (iv) $x \notin y \wedge x \in y'$, and (v) $x \in y \wedge x \notin y'$ are the usual nonantinomic ones. In turn, $x \in y''$, $x \notin y''$, and $x \in \notin y''$ yield corresponding cases for y''. All these cases are of course determined by the kind of membership of x to the relative universe w; that is, if $x \in w$, only cases (ii) and (iv) can obtain; and if $x \in \notin w$, only cases (i) and (iii) are possible. A change in universe may change not only the members x of y', y'', etc., but also the kind of membership of each x to y', y'', etc.

§16. Ordered pairs, relations, functions, cardinalities, Sierpiński's theorem.

Definition 14. (i)(y,z) stands for $\{\{y\},\{y,z\}\}$. The existence of the ordered pair follows Axiom 3; (x,y) could have antinomic members if $\{y\} \in \notin (y,z)$ or $\{y,z\} \in \notin (y,z)$.

(ii) $y \times z$ stands for $\{(u,v)\colon u \in y \wedge v \in z\}$. The existence of the Cartesian product follows Axiom 2; again, $y \times z$ could have antinomic members.

(iii) Rel(R) iff $R \subseteq y \times z$ for some y and z members or subsets of w. R is a binary relation, and if $y=z$, R is a relation on y. R is antinomic iff there exists a pair (u,v) such that $(u,v) \in \notin R$, nonantinomic otherwise. We shall only consider binary relations here.

(iv) If R is a relation, Dom $R=\{u\colon(u,v) \in R\}$, Rng $R=\{v\colon(u,v) \in R\}$. Domain and range of a binary relation.

(v) Given the sets y and z in w, Func(F) iff $F \subseteq y \times z \wedge$ Dom $F=y \wedge$ Rng $F \subseteq z \Rightarrow ((u,v) \in F \wedge (u,t) \in F \Rightarrow v=t)$. F is a function on y into z. F is antinomic iff it is an antinomic relation, nonantinomic otherwise.

(vi) Inj(F) iff $F \subseteq y \times z \wedge$ Func$(F) \wedge ((u,v) \in F \Rightarrow \exists!u((u,v) \in F))$. F is a one-to-one or injective function on y into z. $\exists!u$ is defined by $\exists u \mathrm{A}(u) \wedge \forall r \forall s(\mathrm{A}(r) \wedge \mathrm{A}(s) \Rightarrow r=s)$.

(vii) Sur(F) iff $F \subseteq y \times z \wedge$ Func$(F) \wedge$ Rng $F=z$. F is a function on y onto z, or surjective.

(viii) Bij(F) iff $F \subseteq y \times z \wedge$ Inj$(F) \wedge$ Sur(F). F is a one-to-one function on y onto z, or bijective.

(ix) Two sets y and z have the same cardinality (or are equinumerous) iff $\exists F$(Bij$(F) \wedge$Dom $F=y \wedge$Rng $F=z)$, denoted by Card y=Card z. A set x is inductive, Ind(x), iff $\exists y \exists F$(Nat$(y) \wedge$Bij$(F) \wedge$Dom $F=x \wedge$Rng $F=y)$.

(x) A set y is of cardinality less than or equal to that of z iff $\exists F(F\subseteq y\times z\wedge \text{Inj}(F)\wedge$ Dom $F=y\wedge$Rng $F\subseteq z)$, denoted Card $y\leq$Card z. Card $y<$Card z stands for Card $y\leq$Card $z\wedge$ Card $y\neq$Card z. A set y is reflexive, Ref(y), iff $\exists x\exists F(x\subset y\wedge \text{Bij}(F)\wedge \text{Dom}F=y\wedge \text{Rng}F=x)$.

The equivalence relation Card y=Card z is antinomic iff $\exists F_1(\text{Bij}(F_1)\wedge$Dom $F_1=y\wedge$Rng $F_1=z\wedge\exists(u.v)((u,v)\in\notin F_1))\vee\exists F_2(\text{Bij}(F_2)\wedge$Dom $F_2=z\wedge$ Rng $F_2=y\wedge\exists(u,v)((u,v)\in\notin F_2))$, which allows for three possibilities: (i) Card y_a=Card z iff the first disjunct is true, but the second one is not in the disjunction just given; similarly, (ii) Card $y=_a$ Card z iff there is no F_1 but there is a F_2 for the same formula, and (iii) Card $y_a=_a$ Card z iff there is both a F_1 and a F_2.

Correspondingly, Card $y_a\leq$Card z iff $\exists F(\text{Inj}(F)\wedge$ Dom $F=y\wedge$ Rng $F\subseteq z\wedge$ $\exists(u,v)((u,v)\in\notin F))$. The meaning of Card $y_a<$Card z is obvious.

Definition 15. The cardinal number of a set y, denoted Card y, is the equivalence class of all sets u equinumerous to y in a universe w (Card u=Card y). Card y is a subset of w and hence relative to the given universe: from universe w_1 to universe w_2, Card y may change its members, and y its relative cardinality vis-a-vis other sets. Card y may contain antinomic members as well as members which are and are not equinumerous to y.

In 1947 Sierpiński showed that given a function F on y into z, it is not possible to prove without the axiom of choice that the cardinality of the range of F is not greater than the cardinality of the domain of F. That is, without the axiom of choice, Card $y<$ Card Rng F is not inconsistent with set theory. Clearly, with the antinomic comparability of cardinalities if $\forall F((F\subseteq y\times z\wedge$ Func(F)$\wedge$ Dom $F=y\wedge$Rng $F=z\wedge$Bij(F))$\Rightarrow\exists(u,v)((u,v)\in\notin F))$, then Card $z>$ Card y as well as Card $y\geq$ Card z. Thus, even with the axiom of choice Card $z>$ Card y is not excluded.

§17. Mediate sets.

It was Bolzano in his *Paradoxes of the Infinite* who first distinguished between a set being finite if (i) it is inductive, i.e., counted by a terminal sequence of positive integers, or (ii) it is not reflexive, i.e., equinumerous to a proper subset of itself (today a nonreflexive set is also called Dedekind finite). A mediate set is defined in *Principia Mathematica*[15] as one which is noninductive and nonreflexive. The existence of such sets is ruled out by the axiom of choice; without the axiom of choice, their existence is possible. The cardinality of a mediate set μ is comparable to that of an inductive set x in the sense that Card $x<$ Card μ

(mediate sets contain finite subsets), but it is not comparable to the cardinality of a reflexive (Dedekind infinite) set; that is, $\neg(\text{Card } \mu < \aleph_0) \wedge \neg(\text{Card } \mu \geq \aleph_0)$, a mediate cardinal being the cardinal number of a mediate set. There is neither a minimum nor a maximum mediate cardinal; also, Card $\mu \neq$ Card μ +1 and Card $\mu \neq$ Card μ -1. The mediate cardinals are closed under addition and under multiplication by a mediate cardinal or by an inductive cardinal different from zero. Further, if Card $\mu^{\text{Card } \upsilon}$ is mediate, then μ or υ is mediate. However, if μ is mediate, then $2^{2^{\text{Card } \mu}}$ is not mediate but reflexive (the power set of the power set of a nonempty, noninductive, and nonreflexive set is reflexive). As for $2^{\text{Card } \mu}$ with μ mediate, sometimes it is mediate, sometimes it is reflexive.[15]

A paper by Dorothy Wrinch[16] generalizes the notion of mediate cardinals to those which are comparable to all the usual cardinals up to an aleph greater than or equal to $\aleph_0$. The negation of the existence of such generalized mediate cardinals implies the axiom of choice[16] and is therefore equivalent to such axiom, since the latter implies the nonexistence of all mediate cardinals. Axiom 7 above asserts the existence of nonmediate infinite sets but leaves open the possible existence of mediate sets. One of the various axioms of choice to be proposed here will be relativized to nonmediate sets; yet, choice and mediate sets will be compatible.

Since classically mediate cardinals do not satisfy the axiom of choice, they need not necessarily be comparable. The existence of incomparable mediate cardinals is still an open question. Mediate cardinals have been described as "small" in that they share with inductive sets the property of being nonreflexive, and as "large" because they cannot be obtained by adding 1 to 0 a finite number of times ("finite" used in the intuitive sense that such addition has an effective end). Here, because functions can be antinomic, a set y can be mediate and nonmediate if every function F that maps y onto a proper subset of y contains a pair (u,v) which belongs and does not belong to F. As a consequence, in such a case, $\forall z(z \subset y \wedge$ Card y=Card $z \Rightarrow$ Card $z <$ Card $y)$. A mediate set y which is not nonmediate shall be called *strictly mediate*; if y is both mediate and nonmediate it shall be called *antinomically mediate*. A set can be simultaneously antinomically mediate and the antinomic member of another set. Note that a reflexive set y may have an injective image in an antinomically mediate set z: since z is reflexive and nonreflexive, then y may be comparable and noncomparable to z. In fact, if every function F that compares y to z is not only antinomic and injective but bijective as well, and such that $\forall u \forall v((u,v) \in y \times z \wedge (u,v) \in F) \Rightarrow (u,v) \notin F)$, then z is both mediate and equinumerous to a reflexive set. One should keep in mind, though, that being finite, infinite,

or mediate in any sense are properties relative to the universe w. Changing the universe may make a set reflexive, if it was not, by adding the appropriate function, or make it antinomically reflexive if it was simply reflexive, etc. It is a prejudice to think that mediate sets are useless; like the generic sets produced by forcing methods, they throw light on the understanding of sets in general and on the axiom of choice in particular. More about this later.

§18. Amorphous sets.

The standard definition of infinite set is that of a set not equinumerous with a natural number, and finite if it is. A set is called amorphous if it is infinite in the standard sense but it is not the union of two disjoint infinite sets. There are models of set theory in which the axiom of choice fails and which have amorphous sets; one such is the basic Fraenkel model.

A set y is called Tarski finite (T-finite) iff every nonempty $\subseteq$-monotone chain $X \subseteq \mathscr{P}(y)$ has a $\subseteq$-maximal element. Every amorphous set is T-finite; hence, not every T-finite set is finite in the standard sense. Further, every T-finite set is nonreflexive (Dedekind finite) but the converse is not true.

Since amorphous sets are infinite in the standard sense they are noninductive, and because they are also nonreflexive, they are mediate. One must remember that if we do not assume the axiom of choice, there are several nonequivalent ways of defining infinite sets, as well as finite sets. Thus, a set may be nonfinite in one sense and finite in another. According to Von Neumann, this situation raises serious objections to constructive philosophies of mathematics — intuitionism and the like.[17] The fact is that without the axiom of choice we do not know exactly what finite means, the one concept that constructivism deems fundamental and unmistakable. We must face this issue: without the axiom of choice the idea of finite becomes ambiguous and hazy, and a set can be finite in one sense and infinite in another, as well as being both finite and infinite in the same sense, as is the case with an antinomically mediate set, both Dedekind finite and Dedekind infinite.

§19. An antinomic set theory AS_2 based on inclusion.

In a previous paper[18] we took inclusion instead of membership as the one basic primitive set-theoretic predicate; the other primitive ideas were those of set (x, y, z, ..., variable sets; a, b, a_1, a_2, a_3,..., constant sets), and binary relations (R, S, T, ...). The definitions and axioms offered there are as follows.

Definition 1. $y=z$ iff $\forall x(x\subseteq y\Leftrightarrow x\subseteq z)$. Equality.

Axiom 1. $\forall y\forall z(y=z\Rightarrow\forall u(y\subseteq u\Leftrightarrow z\subseteq u))$. Extensionality.

Definition 2. $y\subset z$ iff $y\subseteq z\wedge y\neq z$. Proper inclusion.

Axiom 2. $\exists y\forall z(z\notin y\wedge\forall u(u\neq y\Rightarrow y\subseteq u))$. Null set, denoted by $\varnothing$ and not included in itself.

Axiom 3. Reflexivity (for all sets other than $\varnothing$), antisymmetry, and transitivity of inclusion.

Axiom 4. $\forall y\exists z((z\subseteq y\wedge\forall x(x\subseteq y\wedge\phi(x)\Rightarrow x\subseteq z)\wedge\forall u\forall v((v\subseteq y\wedge\phi(v)\Rightarrow v\subseteq u)\Rightarrow z\subseteq u))$. Separation, where $\phi(x)$ is any well-formed formula in the language of AS_2 in which y, z, u, and v do not occur and x is a free variable. If $\phi(y)$ is also the case, then $y\subseteq z$, i.e., $y=z$ by antisymmetry. Since some subsets of y may not satisfy ϕ, "separation" does not have the clear-cut meaning that it has in Zermelo's set theory, i.e., it is possible for z to have as subsets sets without the property ϕ.

Definition 3. The notation $z=\{x:x\subseteq y\wedge\phi(x)\}$ represents the least set u that contains all the sets included in y having the property ϕ.

Axiom 5. $\forall x\exists y\exists z(x\subseteq y\wedge z\subseteq y\wedge z\notin x\wedge x\notin z)$. Expansion. There is no class of all sets.

Now let a_1 be an arbitrary but fixed set, and a_2 a nonspecified but fixed superset of a_1 satisfying the condition that y satisfies in Axiom 5, that is, $a_1\subseteq a_2\wedge\exists z(z\subseteq a_2\wedge z\notin a_1\wedge a_1\notin z)$; the existence of this a_2 is guaranteed by the axiom. Let a_3 be a nonspecified but fixed superset of a_2 satisfying the same condition. In general, let a_{n+1} be a similar superset of a_n. The finite sequence a_1, a_2, ..., a_n, a_{n+1} can be made as long as one wishes by successive application of Expansion. However, in order to assert the existence of an infinite set that contains as subsets all the possible terms of this sequence, we need the following additional axiom scheme.

Axiom 6. For any sequence a_1, a_2, ..., a_n, ... satisfying the description just given $\exists y(a_1\subseteq y\wedge(a_n\subseteq y\Rightarrow a_{n+1}\subseteq y))$. Infinity. The infinite set y contains all the sets a_i of the sequence, plus all the subsets of each of these terms.

Axiom 7. $\forall x\forall y\exists z\forall u(u\subseteq x\vee u\subseteq y\Leftrightarrow u\subseteq z)$. Union.

Axiom 8. $\forall x\forall y\exists z\forall u(u\subseteq x\wedge u\subseteq y\Leftrightarrow u\subseteq z)$. Intersection. Union and intersection as determined by these axioms differ from their usual definitions as operations given in terms of membership; for example, no new subsets can be obtained in z by the union of x and y other than those already in x and y.

Definition 4. $E(x)$ stands for $x\neq\varnothing\wedge\forall y(y\neq\varnothing\Rightarrow y\subsetneq x\vee y=x)$. Elementhood. Elements are nonempty sets without nonempty proper subsets. The null set is not an element, although it can be the term of a predicate formula and is certainly the subset of every set except itself.

Schröder asserted that "nothing" is a subject of every predicate, to which Frege objected, drawing the contradictory conclusion that if so, then $\phi(\varnothing)\wedge\neg\phi(\varnothing)$ would obtain, suggesting that if one must have the null set at all, it is better to have it as a subset of every set.[19] From an antinomic point of view both positions can be made simultaneously acceptable.

Axiom 9. $\forall x(x\neq\varnothing\Rightarrow\exists y(y\subseteq x\wedge E(y))$. Regularity. Every set contains at least one element.

Axiom 10. $\forall x\exists y\exists z(x\subseteq y\wedge z\subseteq y\wedge z\not\subseteq x\wedge E(z))$. Element expansion. There is no set of all elements, and there is an infinity of them.

Definition 5. $x\in y$ stands for $x\subseteq y\wedge E(x)$. Membership. Only elements are members. Also, every element is a member of itself, and given two distinct elements, neither one is a member of the other.

Axiom 11. $\exists y\forall x(x\in y\Leftrightarrow\phi(x))$. Comprehension for elements. $\phi(x)$ is a well-formed formula in which y does not occur and x is a free variable. This axiom asserts the existence of a set containing as members all the elements that have the property ϕ. The set of all elements which are not members of themselves is empty, i.e., Russell's paradox cannot be transferred to AS_2.

The objective of the approach just described is to have a set-theoretic base on which to build a topology of multiple location.[18] Here, we shall outline briefly how to use inclusion as an antinomic predicate. Assume that some sets are antinomic in the sense that they are included and not included in another set, that is, $\nvDash x\subseteq y\wedge x\not\subseteq y$ and $\vDash x\subseteq y$, regardless of whether $\vDash x\not\subseteq y$ or not-$\vDash x\not\subseteq y$, abbreviated $x\subseteq\not\subseteq y$, which reads "$x$ is an antinomic subset of y" or "y is an antinomic superset of x." The set y need not be an antinomic subset of another set. In fact, (i) some sets can be antinomic subsets of other sets, $x\subseteq\not\subseteq x$ included as a possibility, and (ii)

some sets may not be antinomic subsets of any set. Symmetrically, (iii) some sets may have only antinomic subsets, (iv) others may have some antinomic subsets and some nonantinomic ones, and (v) some sets, finally, may not have a single antinomic subset.

Similar to the notation proposed for membership, given the constants a and b representing fixed sets, $a\subseteq b$ stands for $\models a\subseteq b$ but not-$\models a\subseteq b$ and not-$\vDash a\subseteq b$, regardless of whether $\vDash a\not\subseteq b$ or not-$\vDash a\not\subseteq b$. Also, $a\not\subseteq b$ stands for $\models a\not\subseteq b$ but not-$\models a\subseteq b$ and not-$\vDash a\not\subseteq b$, regardless of whether $\vDash a\subseteq b$ or not-$\vDash a\subseteq b$. Assuming completeness as we did with AS_1, $\models a\subseteq b$ is metamathematically equivalent to $\vdash a\subseteq b$, which also holds for every positive or negative formula of AS_2, whose axioms can now be extended to include antinomic cases. Thus, for example, $y=z\wedge y\neq z$ may obtain, and the existence of $\mathcal{P}_a y=\{x:x\subseteq\not\subseteq y\}$ be justified by Separation, even though not every set in $\mathcal{P}_a y$ must be an antinomic subset of y. The notions of ordered pair, Cartesian product, relation, function, equinumerosity, and comparability of cardinals given in the earlier paper[18] can also be antinomically extended and an antinomic topology based on inclusion developed. Further, as with membership, the kind of inclusion in $x\subseteq y$ may be relativized to a universe w, and modified from universe to universe. Incidentally, the null set can also be the antinomic subset of other sets.

It should be remarked that Frege, following Schröder, considered inclusion as "the most important relation between sets,"[19] fully identifying the part-whole relation with set-theoretic inclusion. On the other hand, Hao Wang observed that an unavoidable conclusion of the independence of the continuum hypothesis is that, from the point of view of classical set theory, we still do not know what being a subset really means.

§20. An antinomic set theory AS_3 based on union taken as a primitive predicate.

In a previous paper[20]union was used as a primitive binary predicate rather than as an operator. Here we shall expand the predicate of union and make it antinomic. Let us assume a universe of sets $x, y, z, u, v, s, t, x_1, x_2, x_3, \ldots$ in which for some sets x, y, $\cup xy$ holds ("x is united to y"), for other sets u, v, $\neg\cup uv$ holds ("u is disunited from v") and for still other sets s, t, $\cup st\wedge\neg\cup st$ holds ("s is united to and disunited from t").

As with $\in$, $\cup$ is an ambiguous notation to cover both the nonantinomic and the antinomic cases. Accordingly, we shall identify the following possibilities: (i) $\cup\neg\cup xy$ represents the case $\models\cup xy\wedge\neg\cup xy$ and $\vDash\cup xy$, regardless of whether $\vDash\neg\cup xy$ or not-$\vDash\neg\cup xy$; (ii)

$\cup xy$ for the case $\models\cup xy$ but not-$\models\neg\cup xy$ and not-$\mathrm{E}\cup xy$, whether $\mathrm{E}\neg\cup xy$ or not-$\mathrm{E}\neg\cup xy$; (iii) $\neg\cup xy$ for the case $\models\neg\cup xy$ but not-$\models\cup xy$ and not-$\mathrm{E}\neg\cup xy$, whether $\mathrm{E}\cup xy$ or not-$\mathrm{E}\cup xy$.

Axiom 1. $\exists x\forall y(\neg\cup xy)$. There is at least one isolated set strictly disunited from all other sets including itself. If x is one such set, we write $\mathrm{Iso}(x)$.

Axiom 2. $\forall x(\neg\mathrm{Iso}(x)\Rightarrow\cup xx)$; $\forall x\forall y(\cup xy\Rightarrow\cup yx)$. (Union is not necessarily transitive.)

Axiom 3. $\forall y\exists x(\neg\mathrm{Iso}(x)\wedge\neg\cup xy)$. Unity of sets is not universal for nonisolated sets.

Definition 1. $y=z$ iff $\forall x(\cup xy\Leftrightarrow\cup xz)$. In particular: $y\mathrel{=\!\!=}z$ iff $\forall x(\cup xy\Leftrightarrow\cup xz)\wedge\forall u(\cup\neg\cup uy\Leftrightarrow\cup\neg\cup uz)$.

Definition 2. $y\subseteq z$ iff $\forall x(\cup xy\Rightarrow\cup xz)$. Note that $y\subseteq z$ is compatible with $\exists u(\cup\neg\cup uy\wedge\cup uz)$.

Axiom 4. $\forall y\exists z\forall x((\cup xz\Leftrightarrow\cup xy\wedge\mathrm{A}(x))\wedge(\cup\neg\cup xz\Leftrightarrow\cup\neg\cup xy\wedge\mathrm{A}(x)))$, where $\mathrm{A}(x)$ is a well-formed formula in the language of AS_3 in which y and z do not occur and x is a free variable. Separation scheme. Since z is uniquely determined, the notation $z=\{x:\cup xy\wedge\mathrm{A}(x)\}$ is justified. Note that if $\mathrm{Iso}(u)\wedge\mathrm{A}(u)$ is the case, still $\neg\cup uz$ obtains: z does not gather isolated sets.

Axiom 5. For any positive integer k, $\exists x_1, \exists x_2, \ldots, \exists x_k\ (x_1\neq x_2\wedge x_1\neq x_3\wedge\ldots\wedge x_{k-1}\neq x_k)$. This scheme guarantees the existence of an infinity of sets.

Axiom 6. $\forall x\forall y\exists!z(\cup xz\wedge\cup yz\wedge\forall u(\cup uz\Leftrightarrow u=x\vee u=y))$. Paring. The notations $\{x,y\}$ and $\{x\}$ are now justified if we define $\exists!x$, "there exists one and only one x such that ...," by $\exists x\mathrm{A}(x)\wedge\forall u\forall v(\mathrm{A}(u)\wedge\mathrm{A}(v)\Rightarrow u\mathrel{=\!\!=}v)$.

Definition 3. (x, y) stands for $\{\{x\}, \{x, y\}\}$. Ordered pair.

Axiom 7. $\forall u\forall v\exists!z\forall x\forall y(\cup xu\wedge\cup yv\Leftrightarrow\cup(x, y)z)$. Binary Cartesian product denoted by $u\times v$.

Definition 4. (i) R is a binary relation in $u\times v$ means $R\subseteq u\times v$.

(ii) Given $R\subseteq u\times v$, Dom R = $\{x: \cup xu\wedge\cup(x, y)\ R\}$ and Rng $R=\{y:\cup yu\wedge\cup(x, y)\ R\}$. Domain and range of a relation.

(iii)R^{-1} is the inverse relation of R iff $R\subseteq u\times v\wedge R^{-1}\subseteq v\times u\wedge\forall x\forall y\ (\cup(x, y)R\Leftrightarrow\cup(y, x)R^{-1})$.

(iv) F is a function in $u\times v$ iff F is a relation in $u\times v$ and $\forall x\forall y\forall z(\cup(x, y)F\wedge \cup(x, z)F \Rightarrow y{=}z)$.

(v) F is a bijection on Dom F onto Rng F iff F and F^{-1} are both functions.

(vi) Card u = Card v iff there exists a function F which is a bijection on u onto v with $u{=}$ Dom $F \wedge v{=}$Rng F.

(vii) If $z{=}$Dom R = Rng R with R a relation in $u\times v$, then R is a linear ordering on z iff $(\cup xz\Rightarrow\cup(x,x)\ R)\ \wedge(\cup(x,y)R\ \wedge\ \cup(y,x)R\ \Rightarrow\ x{=}y)\wedge(\cup(x,y)R\wedge\cup(y,z)R\ \Rightarrow\ \cup(x,z)R)\wedge\forall x\forall y(\cup xz\wedge\cup yz\Rightarrow\cup(x,y)\mathrm{R}\ \vee\ \cup(y,x)\mathrm{R})$. In particular, R is a well-ordering on z, denoted $R\mathrm{WO}z$ or $\mathrm{WO}(z)$ if R is tacitly assumed to exist, iff R is a linear ordering on z and, in addition, $\forall w(w\subseteq z\wedge\neg\mathrm{Iso}(w) \Rightarrow \exists!s(\cup sw\wedge\forall t(\cup tw\Rightarrow\cup(s,t)R)))$.

It should be mentioned that instead of union, intersection can be taken as an antinomic predicate, antinomic sets being those that satisfy $\cap xy\wedge\neg\cap xy$. This will not be pursued here.

IV. Antinomic Axioms of Choice

E. Hobson proved in 1905 that the standard axiom of choice does not rule out the existence of antinomic sets.[21] Obviously, an antinomic axiom of choice should be based on such sets. More recently, it has been shown that the standard axiom of choice implies the law of excluded middle.[22] The proof, however, breaks down if one assumes the logic outlined in this paper; thus, the antinomic versions of the axiom of choice to be proposed will not imply excluded middle, although they will be compatible with specific instances of this law. (The proof that standard choice implies excluded middle uses contradiction, which is why it fails here.) More important, antinomic versions of the axiom of choice are compatible with sequences of more than two alternatives: $\phi_1\vee\neg\phi_1\vee\phi_2\vee\phi_3\vee\ldots\vee\phi_n$. To bring antinomicity to choice, then, is in keeping with the fact that although $\phi\vee\neg\phi$ understood as an exclusive alternative simplifies logic, the situation in mathematics and the natural sciences is replete with instances in which ϕ and $\neg\phi$ are far from being the only options available.

§21. Antinomic axioms of choice for AS_1.

Because the standard proof of the equivalence of the axiom of choice with, say, the well-ordering principle relies on contradiction, we cannot assume here that the axiom implies

the principle or vice versa — as is the case with most equivalent forms of the axiom of choice. This nonequivalence has, in effect, its positive consequences in that it returns to each of these forms some of the independence, strength, and breadth of scope with which they were originally conceived.

Well ordering can be defined as follows within AS_1.

Definition 1. WO(z) or equivalently RWOz iff $\exists R(\mathrm{Rel}(R) \wedge \mathrm{Dom}\ R = \mathrm{Rng}\ R = z \wedge \forall x(x \in z \Rightarrow (x,x) \in R) \wedge \forall x \forall y((x,y) \in R \wedge (y,x) \in R \Rightarrow x = y) \wedge \forall x \forall y \forall u((x,y) \in R \wedge (y,u) \in R \Rightarrow (x,u) \in R) \wedge \forall x \forall y(x \in z \wedge y \in z \Rightarrow (x,y) \in R \vee (y,x) \in R) \wedge \forall v(v \subseteq z \wedge v \neq \varnothing \Rightarrow \exists s(s \in z \wedge \forall t(t \in z \Rightarrow (s,t) \in R)))$. The expression **WO**($z$) means $=$WO(z) but not-$=\neg$WO(z) and not-E WO(z), whether E$\neg$WO(z) or not-E$\neg$WO(z); $\neg$**WO**(z) means $=\neg$WO(z) but not-$=$WO(z) and not-E$\neg$WO(z), whether EWO(z) or not-EWO(z); WO$\neg$WO(z) means $=$WO(z)$\wedge\neg$WO(z) and EWO(z), whether E$\neg$WO(z) or not-E$\neg$WO(z). **WO**(z) will read "z is strictly well-ordered," WO$\neg$WO(z) will read "z is antinomically well-ordered."

Similarly, the predicate Med(z), "z is mediate," can be defined in AS_1 as $\neg$Ind(z)$\wedge\neg$Ref(z), using the definitions of Ind(z), "z is inductive," and Ref(z), "z is reflexive," given in §16. **Med**(z), $\neg$**Med**(z), and Med$\neg$Med(z) stand for "z is strictly mediate," "z is strictly nonmediate," and "z is antinomically mediate," defined respectively as above for **WO**(z), and $\neg$**WO**(z), and WO$\neg$WO(z).

An antinomic axiom of choice AAC for AS_1 may be introduced in a number of ways; we shall select two of them. The idea is that AAC should not apply to all sets but only to those which are, for example, well-ordered or nonmediate. That is, we shall break the universe w into two classes not necessarily disjoint; in one case, the class of well-ordered sets and the class of non-well-ordered sets; in the second case, the class of nonmediate sets and the class of mediate sets. Accordingly, we have the following two axioms in which $\mathscr{F}$ is a given family of sets, A is a member of the family $\mathscr{F}$, and $\mathscr{C}$ is the choice set.

Axiom 1. $\mathrm{WO}(\mathscr{F}) \Rightarrow \exists \mathscr{C}_1 \forall x(x \in \mathscr{C}_1 \Leftrightarrow \exists A(A \in \mathscr{F} \wedge x \in A \wedge \forall y(y \in \mathscr{C}_1 \wedge y \in A \Rightarrow x = y)))$. Choice for well-ordered sets.

Axiom 2. $\neg\mathrm{Med}(\mathscr{F}) \Rightarrow \exists \mathscr{C}_2 \forall x(x \in \mathscr{C}_2 \Leftrightarrow \exists A(A \in \mathscr{F} \wedge x \in A \wedge \forall y(y \in \mathscr{C}_2 \wedge y \in A \Rightarrow x = y)))$. Choice for nonmediate sets.

Both axioms leave room for subuniverses in which AAC does not apply; both are also ambiguous in the sense that all the following are possible: (i) **WO**($\mathscr{F}$) or WO$\neg$WO($\mathscr{F}$), (ii) $x\in \mathscr{C}_1$ or $x\in\notin \mathscr{C}_1$, (iii) $A\in \mathscr{F}$ or $A\in\notin \mathscr{F}$, (iv) $x\in A$ or $x\in\notin A$, (v) $\neg$**Med**($\mathscr{F}$) or Med$\neg$Med($\mathscr{F}$), and (vi) $x\in \mathscr{C}_2$ or $x\in\notin \mathscr{C}_2$. Classically, (i) WO($\mathscr{F}$)$\Rightarrow$WO($\mathscr{C}_1$) and (ii) $\neg$Med($\mathscr{F}$) $\Rightarrow\neg$Med($\mathscr{C}_2$), since the injective correspondence given by $F{:}A\rightarrow x$ mapping $\mathscr{F}$ onto $\mathscr{C}_1$ (or $\mathscr{C}_2$), with x the unique representative of A in $\mathscr{C}_1$ (or $\mathscr{C}_2$), makes $\mathscr{C}_1$ well-ordered (and $\mathscr{C}_2$ nonmediate). We shall assume implications (i) and (ii); hence, **WO**($\mathscr{F}$) and $\neg$**WO**($\mathscr{C}_1$) will be incompatible (and so will $\neg$**Med**($\mathscr{F}$) and **Med**($\mathscr{C}_2$)).

Axioms 1 and 2 are ambiguous in that the kind of membership of x to $\mathscr{C}_1$ (or to $\mathscr{C}_2$) is not uniquely determined, making $\mathscr{C}_1$ (and $\mathscr{C}_2$) not uniquely determined either. The situation is similar to the one which arose in connection with the separation axiom. Therefore, in order to determine the choice sets $\mathscr{C}_1$ and $\mathscr{C}_2$ more specifically, we must state the following. Note that x and A, either one or both, can be well-ordered (or nonmediate), or the opposite.

Axiom 1′. $\mathrm{WO}(\mathscr{F})\Rightarrow\exists\mathscr{C}_1'(\forall x(x\in \mathscr{C}_1'\Leftrightarrow\exists A(A\in\mathscr{F}\wedge x\in A\wedge\forall y(y\in \mathscr{C}_1'\wedge y\in A\Rightarrow x=y))\wedge \forall u(u\in\notin \mathscr{C}_1'\Leftrightarrow\exists A(A\in\mathscr{F}\wedge u\in\notin A\wedge\forall v(v\in\notin \mathscr{C}_1'\wedge v\in\notin A\Rightarrow u=v))\wedge\forall r\forall s(r\in \mathscr{C}_1'\wedge r\in A\wedge s\in \mathscr{C}_1'\wedge s\in A \Rightarrow r=s))$.

Axiom 2′. With $\neg$Med($\mathscr{F}$) as a premise, same conclusion as in Axiom 1′ replacing $\mathscr{C}_1'$ by $\mathscr{C}_2'$.

Since in the four preceding axioms the major implication goes only in one direction, the existence of a choice set does not mean that $\mathscr{F}$ must be well-ordered or nonmediate; in fact, $\mathscr{C}_1'$ and $\mathscr{C}_2'$ themselves can be non-well-ordered or mediate respectively. If $\mathscr{F}$ is strictly mediate, the bijection $F{:}A\rightarrow x$ on $\mathscr{F}$ onto $\mathscr{C}_2'$ makes the latter strictly mediate, as the following metamathematical reasoning shows. If $\mathscr{C}_2'$ were antinomically mediate, then a bijection G would exist that maps $\mathscr{C}_2'$ onto a proper subset of itself y; however, such mapping must have at least a pair (u,v) such that $u\in \mathscr{C}_2'$, $v\in y$, and $(u,v)\in\notin G$, since $\mathscr{C}_2'$ is both reflexive and nonreflexive. The composition of the three mappings F, G, and F^{-1} in this order is a bijection on $\mathscr{F}$ onto a proper subset of itself, where F^{-1} is the inverse of F. But then $\mathscr{F}$ would be both strictly mediate and antinomically mediate, which is a metamathematical impossibility. Of course, as mentioned above, $\mathscr{F}$ could be both well-ordered and nonmediate, and the bijection $F{:}A\rightarrow x$ in each case would yield the necessary function to also make $\mathscr{C}_1'$ and $\mathscr{C}_2'$ well-ordered and nonmediate, respectively.

Other set-theoretic properties could be used to make room for a subuniverse in which the axiom of choice holds and in whose complement it does not necessarily hold. For example, since using choice it cannot be proved that every set is similar to an ordinal, once the ordinals are introduced one could use Count(x), "x is similar to an ordinal," as a substitute for either WO(x) or ⌉Med(x), which would lead to an alternative version of Axioms 1 and 2. The same applies to other principles usually given as equivalents of the axiom of choice. We shall not pursue this matter here.

Let us finally link well-ordered sets and nonmediate sets with the following.

Definition 2. $w_0=\{x: x\in w \wedge \urcorner \mathrm{Med}(x)\}$, w_0 is the nonmediate subuniverse.

Axiom 3. $\mathrm{WO}(x)\Rightarrow x\in w_0$, every well-ordered set is nonmediate.

§22. Ordinals.

We now represent the sequence $\varnothing$, $\{\varnothing\}$, $\{\varnothing, \{\varnothing\}\}$, ... by 0, 1, 2, ..., the finite ordinals or natural numbers. Using the axiom of infinity, let us call ω the intersection of all sets that have 0,1,2,..., as members. $S\omega=\omega\cup\{\omega\}$ will be denoted by $\omega+1$, etc. We then define the sequence of ordinals *Ord* in the usual way:

Definition 3. $\alpha\in Ord$ iff (i) $x\in y\in\alpha\Rightarrow x\in\alpha$ and (ii) $\forall z(z\in\alpha\Rightarrow(u\in v\in z\Rightarrow u\in z))$.

Ordinals will be represented by Greek letters except for the class of all ordinals *Ord*, which is also an ordinal, and belongs and does not belong to itself. For ordinals α and β we have the following:

Axiom 4. $\forall\alpha\forall\beta(\alpha\in\beta\vee\alpha=\beta\vee\beta\in\alpha)$. In addition, $\alpha\in Ord\Rightarrow\in \mathrm{WO}\alpha$ and $\alpha\in Ord\Rightarrow\urcorner\mathrm{Med}(\alpha)$, every ordinal is well-ordered by $\in$ and no ordinal is mediate. Finally, *Ord* is well-ordered by $\in$.

Essentially, ordinals behave like their standard counterparts, although they can be antinomic members of sets which are not ordinals. Addition, multiplication, and exponentiation of ordinals can be defined inductively in the usual way, and the necessary theorems postulated whenever their classical proofs include negative formulas; such theorems include the principle of transfinite induction, the statement that ω is the smallest limit ordinal, the uniqueness of ordinal operations, etc.

Extending to set theory an idea introduced in a previous paper for formal arithmetic,[14] let us now add strict order as a primitive antinomic predicate.

Axiom 5. $\forall\alpha\forall\beta\forall\gamma((\alpha<\beta\wedge\beta<\gamma\Rightarrow\alpha<\gamma)\wedge(\neg\mathrm{Nat}(\alpha)\wedge\neg\mathrm{Nat}(\beta)\Rightarrow\alpha<\beta\wedge\beta<\alpha)\wedge(\alpha<Ord\wedge Ord<\alpha))$. Transitivity of $<$ in *Ord*, and symmetry (hence reflexivity) of $<$ for all nonfinite ordinals. Each nonfinite ordinal is greater and less than all other ordinals including itself, that is, every geometric representation of $<$ requires bilocation. Whereas the order type of the $\in$-ordering of *Ord* is $\overline{Ord}$, the usual one, the $<$-ordering of *Ord* has the following order type: $1+\overline{Ord}^{*}+\omega+\overline{Ord}+1$, "1" being the order type of *Ord* itself, the greatest and hence the least ordinal, $\overline{Ord}^{*}$ is the mirror image of $\overline{Ord}$, the last being the standard type of the set of nonfinite ordinals, and ω the order type of the set of finite ordinals placed in the middle of any model. Each finite ordinal has simple location, and each nonfinite ordinal has one location to the right and another to the left of the fragment of all finite ordinals, that is, one in the segment of type $1+\overline{Ord}^{*}$ and another in the segment of type $\overline{Ord}+1$.

§23. A mediate continuum hypothesis.

Although by Axiom 3 every well-ordered set x is nonmediate, as already mentioned x is not necessarily similar to an ordinal; Gödel's indirect proof that every well-ordered set is similar to an ordinal cannot be carried out in AS_1.[23] Further, the converse of Axiom 3 does not hold, as the following classical example shows: if μ is strictly mediate, $\mathscr{P}(\mu)$ is either strictly mediate or strictly nonmediate, and $\mathscr{P}\mathscr{P}(\mu)$ is strictly nonmediate, yet although reflexive, the latter is not well-ordered since it contains as a subset a replica of μ. In other words, some nonmediate sets have mediate subsets, whereas all mediate sets have nonmediate subsets.

The cardinal number Card x of a set x was defined in §16 as the equivalence class of all sets equinumerous to the set x; Card x is a subset of the universe w and is relative to that universe. Each set $x\in w$, then, has a cardinal number Card x regardless of the kind of order it may have, and whether or not x is a mediate set.

The alephs can now be defined as follows.

Definition 4. $\aleph_\alpha$ is the cardinal number of a given nonfinite ordinal γ. The class of all alephs is well-ordered as follows: $\aleph_\alpha\leq\aleph_\beta$ iff $\gamma_1\in\gamma_2\vee\gamma_1=\gamma_2$, where γ_1 and γ_2 are any ordinals such that $\gamma_1\in\aleph_\alpha$ and $\gamma_2\in\aleph_\beta$.

Axiom 6. For every ordinal α there exists a cardinal number $\aleph_\alpha$. The class of all alephs is not only well-ordered but it is also similar to *Ord*. Since mediate sets are members of *w*, not every nonfinite set has an aleph for its cardinal number.

Cardinal arithmetic can be defined as follows. Let us symbolize cardinal numbers with bold face letters **m**, **n**, ..., and let *m*, *n*, ..., be any representative of the classes **m**, **n**, ..., respectively; **1**, **2**, **3**, ..., are Card 1, Card 2, Card 3,....

Definition 5. (i) $\mathbf{m}+\mathbf{n}$ is the cardinal number of the disjoint union of *m* and *n*; (ii) $\mathbf{m}\bullet\mathbf{n}$ is the cardinal number of the Cartesian product $m\times n$; (iii)$\mathbf{m}^{\mathbf{n}}$ is the cardinal number of the set of all functions on *m* into *n*. The antinomicity of some of the entities involved in (i)-(iii) does not affect the uniqueness of the operations defined.

The beth numbers are defined as follows.

Definition 6. $\beth_0=\aleph_0$, $\beth_{\alpha+1}=\mathbf{2}^{\beth_\alpha}$

Assuming the generalized continuum hypothesis (GCH), $\beth_\alpha=\aleph_\alpha$, and the beth notation becomes superfluous. GCH is not assumed here, and the relation between the alephs and the beths is left undetermined. In addition to these two kinds of nonfinite cardinals, now we need to introduce two more, given that not every nonfinite cardinal is an aleph or a beth.

Definition 7. If $\boldsymbol{\mu}$=Card μ is the mediate cardinal of a mediate set μ, then $\mathbf{2}^{\boldsymbol{\mu}}=\gimel_\mu$ is a gimel number indexed by μ to indicate it provenance. Gimel numbers can be mediate or nonmediate, and only some mediate numbers are gimel numbers.

Definition 8. If $\boldsymbol{\mu}$=Card μ is the mediate cardinal of a mediate set μ, then $2^{2^\mu}=\daleth_\mu$ is a daleth number indexed by μ to indicate its provenance. Daleth numbers are nonmediate. Whether $\daleth_\mu$ is an aleph, a beth, or another yet undefined kind of nonfinite reflexive cardinal is left as an open question. The relation between daleths and gimels is given by Axiom 8 below.

Axiom 7. Card $\mu\neq$Card $\mu'\Rightarrow(\gimel_\mu\neq\gimel_{\mu'}\wedge\daleth_\mu\neq\daleth_{\mu'})$.

The gimel and daleth numbers are not linearly ordered, and even if a gimel number is nonmediate, it is not necessarily equal to a daleth, a beth, or an aleph. Further, if the daleths were beth or aleph numbers, they would be well-ordered by the ordinals, thus inducing

a well-ordering of the mediate sets. However, we shall postulate the following mediate continuum hypothesis (MCH).

Axiom 8. $\mathrm{Med}(\mu)\Rightarrow \exists\mu'(\mathrm{Med}(\mu')\wedge \top_{\mu}=\beth_{\mu'})$. Every daleth equals a gimel number, i.e., the cardinal number of the power set of the power set of a mediate set is the cardinal number of the power set of some mediate set.

From the viewpoint of the Foundations of Mathematics, Axiom 2 has the advantage over Axiom 1 of making the choice operation independent of order, for there is indeed something more basic about choice than any kind of ordering that one might attach to a set. But as mentioned, the alternative of taking $\mathscr{F}$ as nonmediate to guarantee the existence of a choice set is not indispensable either: $\mathscr{F}$ could be merely nonamorphous, in which case some mediate families $\mathscr{F}$ could also yield a choice set. However, it seems rather forced to extrapolate the well-ordering principle from the set of natural numbers to all unimaginable sets simply to be able to single out a definite individual from every nonempty set. And it seems just as forced to identify infinity with Dedekind infinity since, for example, it is shortsighted to assume that nonfinite nonreflexive sets are useless because we have not yet found any use for them. In contrast, the operation of choice is itself truly primitive and intuitively natural whenever it is applicable. Although not always feasible, it is essential even for selecting the very first symbol to put on paper. Indeed, choice is as indispensable from a mathematical point of view as the equally primitive operation of comprehension, i.e., the gathering of individuals that share in a given property. Still other antinomic versions of the axiom of choice should yield new foundational approaches as well as new structural understanding of these two fundamental mathematical operations of choosing and gathering.

§24. Axioms of choice for AS_3.

Axioms of choice for AS_2 and AS_3 parallel those proposed for AS_1. Let us look briefly at the case of AS_3.

Axiom 1. $\mathrm{WO}(\mathscr{F})\Rightarrow\exists\mathscr{C}(\forall x(\cup x\mathscr{C}\Leftrightarrow\exists A(\cup A\mathscr{F}\wedge\cup xA\wedge\forall y(\cup y\mathscr{C}\wedge\cup yA\Rightarrow x=y)))\wedge\forall u(\cup\rceil\cup u\mathscr{C}\Leftrightarrow\exists A(\cup\rceil\cup A\mathscr{F}\wedge\cup\rceil\cup uA\wedge\forall v(\cup\rceil\cup v\mathscr{C}\wedge\cup\rceil\cup vA\Rightarrow u=v)))\wedge\forall r\forall s(\cup r\mathscr{C}\wedge\cup rA\wedge\cup s\mathscr{C}\wedge\cup sA\Rightarrow r=s)$. The predicate $\mathrm{WO}(z)$ was defined in §20.

As is the case with AS_1, premises other than $\mathrm{WO}(\mathscr{F})$ may condition the existence of choice set $\mathscr{C}$; for example, we may gather all the sets generated by applications of separation scheme 3 given in §20, as shown in the following definition:

Definition 1. Comp(z) means z exists by virtue of Axiom scheme 3, §20, i.e., there is a well-formed formula $A(x)$ in the language of AS_3 which gathers z. If the language of AS_3 is uncountable, there would be an uncountable number of such formulas, and potentially an uncountable number of sets z satisfying Comp(z).

Axiom 2. With Comp ($\mathscr{F}$) as a premise, same conclusion as in Axiom 1 above. Again, not only is the existence of a choice set not equivalent to the well-ordering of $\mathscr{F}$ but also is not equivalent to the "predicability" of $\mathscr{F}$ as given in the definition of Comp(z) just proposed. (It is ironic that here choice depends on predicability even if it is nonconstructive.)

§25. A final remark.

The logic on which the set theories AS_1, AS_2, AS_3 are based is obviously a limit one in that, apart from its positive fragment, it is built semantically, posing negative formulas true, false, or both when desired, then postulating the true and antinomic ones as axiom-theorems — syntax following semantics except for some metamathematical reasonings. At the level of formulas, this is not unlike the device of adding an uncountable number of constant symbols to the language of a theory in order to use them syntactically in the formation of terms and formulas. These symbols provide a name for each individual in the universe of a given structure, thus producing an uncountable number of formal atomic sentences from which to gather those which are true in the structure. The notion of diagram introduced by A. Robinson employs these constants and is the set of atomic sentences true in the given structure. This diagram constitutes a ready-made complete theory.[24] Here, the structure is not given in advance, and negative formulas are successively incorporated as true or antinomic in the development of what we may call an "open diagram," a progressing diagram that keeps adding determining characteristics and entities to the models of the true and antinomic formulas previously posited. The purpose is not to obtain a syntactically complete theory but to establish the existence of desirable entities or to modify those already introduced. The next step in the evolution of this and other chapters of antinomic mathematics should move from this limit position toward one more proof-theoretically balanced. How far it is possible to go in this direction and how advantageous it would be to do so are open questions. Yet, the effort involved cannot fail to throw valuable light on the foundational problems that have been touched upon here.

Notes

1. "Many of the most profound results in modern logic have arisen from the analysis of the paradoxes." E.W. Beth, *The Foundations of Mathematics*, p. 481, North Holland, Amsterdam, 1958.

2. See *Paraconsistent Logic: Essays on the Inconsistent*, edited by G. Priest, R. Routley, and J. Norman, Philosophia Verlag, Munich, 1989.

3. J. van Heijenoort, *From Frege to Gödel*, p. 114, Harvard University Press, Cambridge, 1967.

4. K. Gödel, "Russell's Mathematical Logic," *The Philosophy of Bertrand Russell*, p. 131, edited by P.A. Schilpp, Tudor, 1944.

5. Ib., p. 139.

6. Ib., p. 140.

7. B. Russell, *An Enquiry into Meaning and Truth*, pp. 99, 174, Allen & Unwin, London, 1940.

8. F.G. Asenjo, "The Idea of a Calculus of Antinomies," La Plata, 1953, and "A Calculus of Antinomies," *Notre Dame Journal of Formal Logic*, VII, 1966., p. 103.

9. E. Mendelson, *Introduction to Mathematical Logic*, Wadsworth, Monterey, 1987, p. 29.

10. This expands the idea of antinomic model introduced in "Logic of Antinomies," F.G. Asenjo and J. Tamburino, 1975, *Notre Dame Journal of Formal Logic*.

11. Cf. Whitehead and Russell, *Principia Mathematica*, 1927, Vol I. p. 30.

12. Cf. T. Jech, *The Axiom of Choice*, North Holland, Amsterdam, 1973, p. 153.

13. Cf. A. Kolmogorov, "On the principle of excluded middle," J. van Heijenoort, op. cit., pp. 431, 436.

14. Cf. F.G. Asenjo, "Toward an Antinomic Mathematics," *Paraconsistent Logic* (see Note 2), p. 407.

15. *Principia Mathematica*, Vol. II, p. 280.

16. Dorothy Wrinch, "On Mediate Cardinals," *American Journal of Mathematics*, 1923, Vol. 45, pp. 87-92.

17. J. von Neumann, "An Axiomatizaton of Set Theory," J. van Heijenoort, loc. cit., pp. 421-3.

18. F.G. Asenjo, "Formalizing Multiple Location," *Non-Classical Logics, Model Theory, and Computability*, edited by A.I. Arruda, N.C.A. da Costa, and R. Chuaqui, North Holland, Amsterdam, 1977, pp. 25-36.

19. G. Frege, "A critical elucidation of some points in E. Schroeder's *Algebra der Logic*," *Translations from the Philosophical Writings of Gottlob Frege*, edited by P. Geach & M. Black, Oxford, 1977, pp. 86-106.

20. F.G. Asenjo, "Continua Without Sets," *Logic and Logical Philosophy*, 1993, Vol. 1, pp. 95-128.

21. G.H. Moore, *Zermelo's Axiom of Choice*, Springer-Verlag, New York, 1982, pp. 128-9.

22. Cf. M.J. Beeson, *Foundations of Constructive Mathematics*, Springer-Verlag, Berlin, 1985, p.163.

23. Cf. H. Rubin & J. Rubin, *Equivalents of the Axiom of Choice*, North Holland, Amsterdam, 1963, p.8.

24. A. Robinson, *Introduction to Model Theory and to the Metamathematics of Algebra*, North Holland, Amsterdam, 1965, p. 24. Robinson's definitions of "positive" and "negative" diagram are different from the ones given for the same expressions in §14 above in connection with Axiom 8.

Rev. Academia de Ciencias. 49 (1994)

HISTORIA DE LA ENSEÑANZA DEL ALGEBRA EN LA UNIVERSIDAD COMPLUTENSE

Concepción Romo Santos
Departamento de Algebra
Universidad Complutense
MADRID

Introducción

En 1993 se cumplen 700 años de la fundación de la Universidad Complutense, una efemérides que debe ser motivo de orgullo y reflexión. Los que hoy formamos parte de ésta Universidad, estamos orgullosos de ser complutenses y vamos a unirnos a los actos conmemorativos de su VII centenario. Estudiaremos la historia de la enseñanza del Algebra en nuestra querida Universidad.

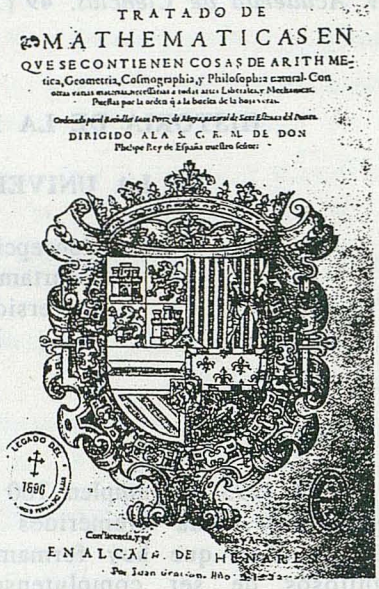
§1.- Historia de la enseñanza de las Matemáticas en la Universidad Complutense hasta su traslado a Madrid.

Sancho IV ordenó que se estableciera el Estudio General de Alcalá de Henares en su Real Carta de 1293.

El historiador Esteban Azaña es quien aporta más datos sobre el sistema educativo alcalaino a lo largo del siglo XIV y durante la primera mitad del siglo XV. Asegura el cronista decimonónico que el Estudio medieval de Alcalá de Henares, aunque débil y sin duda intermitente, tuvo vida académica a lo largo del siglo XIV. Pero, sin embargo, es en el siglo XV donde empezó su esplendor, ya que en 1421 se contaba con cursos de hebreo, MATEMATICAS y música. De mediados del siglo XV son las tres cátedras que dotó el arzobispo Carrillo con los frutos y rentas provenientes de los beneficios de su diócesis. Una era de Gramática, otra de Lógica y la tercera de Ciencias.

La empresa iniciada por Sancho IV y conbtinuada por Carrillo llega a su fin en las postrimerías del siglo XV con la fundación, por el Cardenal Cisneros de San Ildefonso en Alcalá de Henares.

La Complutense Cisneriana se centra en torno al Colegio Mayor de San Ildefonso y abarca el periodo comprendido entre 1499 y 1545.



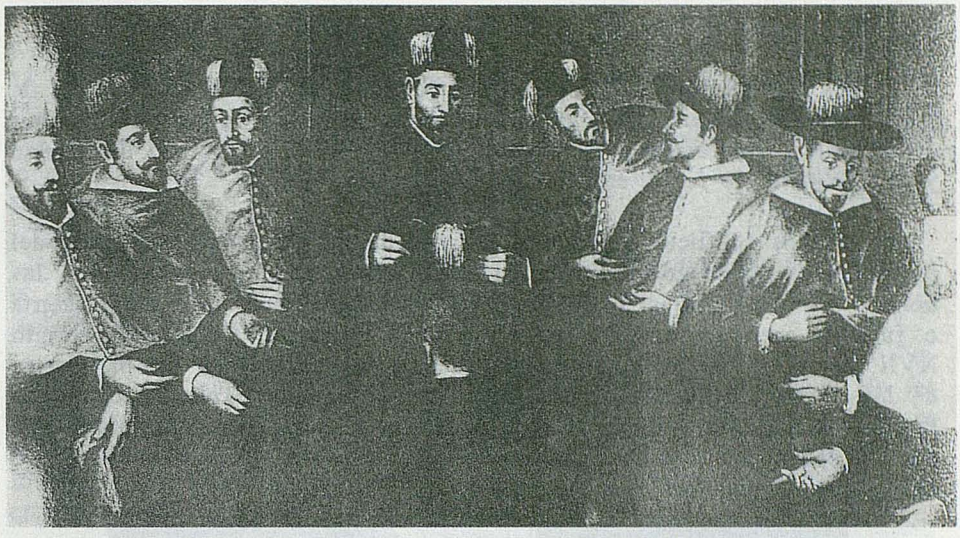
Efigie del Cardenal Cisneros Tratado de matemáticas (siglo XVI)

La Universidad Complutense durante elperiodo de Cisneros contó con cuatro Facultades: Teología, Artes. Medicina y Derecho Canónico y con dos Escuelas de Gramática, que operaban en los Colegios Menores de San Eugenio y San Isidoro.

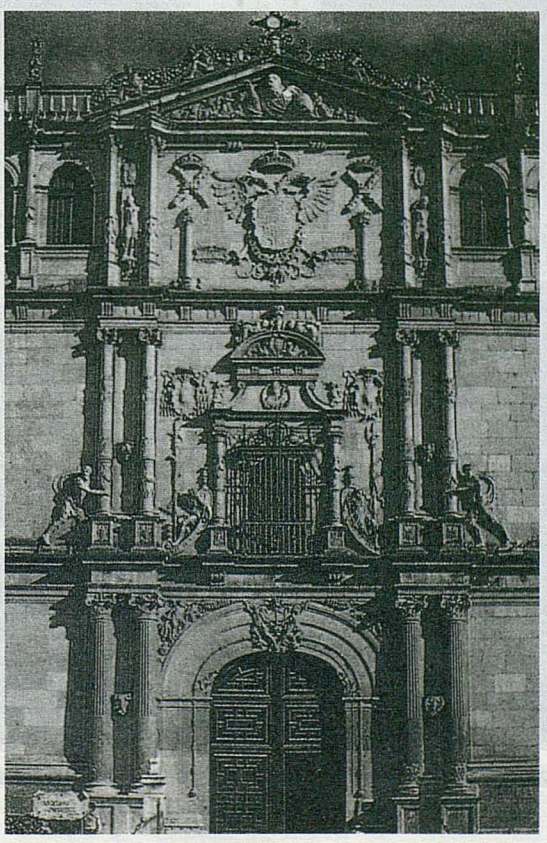
La Facultad de Artes Liberales expedía títulos de bachiller, licenciado y maestro. El Bachillerato y la Literatura se cursaban en cuatro años - dos cada uno - y se estudiaban las siguientes asignaturas: Súmulas logicales, Predicamentos, Hermenéutica, Tópicos, Elencos, MATEMATICAS, GEOMETRIA, PERSPECTIVA, Etica, Filosofía natural y la Metafísica de Aristóteles.

Estudiaremos ahora la historia de la Complutense postcisneriana que dura casi dos siglos y coincide con el periodo de gobierno de los Austrias.

Su punto de partida va unido a la capitalidad de Madrid, decretada por Felipe II en 1561, y a la creación en ésta ciudad, en 1545, de los Estudios de San Isidro, una escuela de enseñanza secundaria destinada a la preparación de los hijos de la nobleza. A partir de esa fecha comienza a decaer, aunque lentamente, la Universidad que creó Cisneros en Alcalá y a consolidarse el Instituto de Madrid.



La investidura del grado de doctor estaba rodeada de gran ceremonial, al que había que asistir adornado con las mejores galas

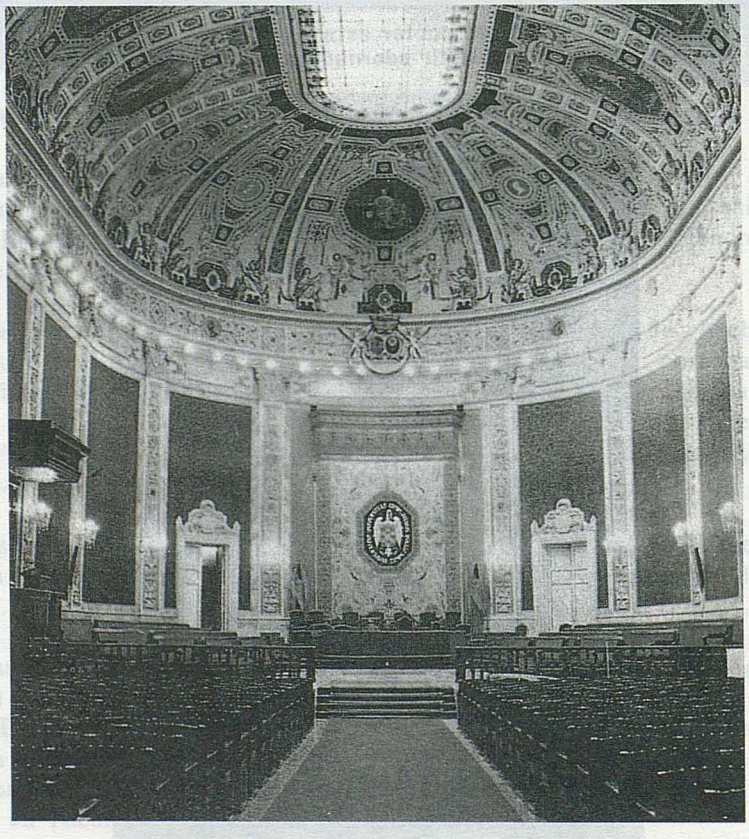


Fachada del Colegio Mayor de San Ildefonso

§2.- La creación de la Universidad Central de Madrid.

La Universidad Complutense postcisneriana, aunque gozó de momentos de esplendor, tuvo que soportar múltiples conflictos: luchas estudiantiles, guerra de Comunidades, litigios con el Arzobispo de Toledo y la Compañía de Jesús, etc.

Dichos conflictos, unidos al despojo de su patrimonio por parte del poder real, al desgobierno de algunos rectores, que conculcaron las normas constitucionales dictadas por el cardenal Cisneros y a la sangría o fuga de cerebros que se produjo en las primeras décadas del siglo XVII, debido a que la Inquisición hizo huir de sus aulas a los partidarios de las doctrinas filosóficas de Erasmo de Rotterdam, explican los varios intentos de traslado de la Complutense a Madrid. que culminarán un siglo después. Su acta de defunción se produce el 29 de octubre de 1836 con la Real Orden que dispone su traslado a Madrid.



Aula Magna de la Universidad de la calle San Bernardo, antes noviciado de los jesuitas

En el periodo que transcurre entre 1836 y 1845 se organizó la Universidad Central de Madrid según el modelo napoleónico. La Facultad de Ciencias quedaría instalada en la capilla de los Reales Estudios. En 1857 se promulgó la Ley de Instrucción Pública - Ley Moyano - en la que se estructuró definitivamente la enseñanza contemporánea hasta la reciente Ley de 1970. En esta ley quedó establecida como Facultad la de Ciencias exactas.

En 1876 se creó la Institución Libre de Enseñanza que propugnó la coeducación y la enseñanza del Arte. Asímismo creó la Junta para la Ampliación de Estudios e Investigaciones Científicas.

En los cursos celebrados entre 1875 y 1902 la matrícula de las asignaturas ascendió de modo progresivo. Hombres brillantes por sus contribuciones científicas fueron José Echegaray, Antonio Aguilar, que instaura en Madrid el Observatorio Astronómico, Carlos Yebes, catedrático de Geometría, y el matemático Gumersindo Vicuña.

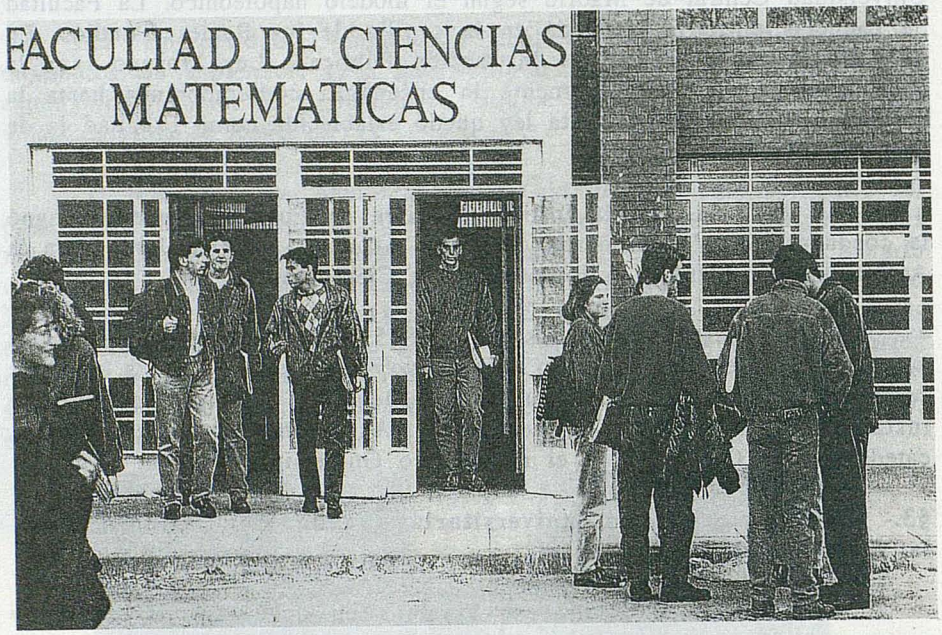
§3.- La nueva ciudad universitaria creada por Alfonso XIII en la finca "La Moncloa".

La Universidad Central se asentaba en una serie de locales y caserones diseminados por el casco antiguo de Madrid, que se encontraban desfasados y adolecían de graves defectos acústicos, lumínicos y térmicos. Alfonso XIII, a sugerencia de su odontólogo, que había estudiado en Estados Unidos, decidió construir la Ciudad Universitaria. Se organizó una junta constructora y se abrió una suscripción popular con 2.500.000 pts. aportadas por la Familia Real. A principios de 1929 comenzaron las obras.

La marcha de las obras fué eficaz e ininterrumpida; el rey presidió la última junta el 5 de abril de 1931. Nueve días después se proclamó la Republica y el rey Alfonso XIII marchaba al exilio. El cambio político fué, sin embargo, beneficioso para el gran proyecto del monarca. Pronto comenzaron las inauguraciones. La Facultad de Filosofía y Letras fué la primera de ellas.

Desgraciadamente, la Ciudad Universitaria se convirtió en primera línea de batalla durante la Guerra Civil y fué duramente castigada por los bombardeos. El resultado fué la destrucción de los fondos documentales y científicos y de los edificios.

Terminada la contienda el arquitecto López Otero continuó tecnicamente al frente de la junta constructora de la Ciudad Universitaria, creada de nuevo en 1940.



Facultad de Ciencias Matemáticas en la Ciudad Universitaria

Complemento de la Universidad habrá de ser el Consejo Superior de Investigaciones Científicas, creado por Decreto Ley de 1939. Su misión era fomentar, orientar y coordinar la investigación científica en España, división CIENCIAS MATEMATICAS y de la Naturaleza.

Entre las Facultades que se crearon en la Ciudad Universitaria estaba la de Ciencias. En el periodo más reciente se ha construido una nueva Facultad de Matemáticas, inaugurada en junio de 1992, siendo rector D. Gustavo Villapalos. Las especialidades que en la actualidad se pueden estudiar en la licenciatura de Ciencias Matemáticas, son las siguientes: Fundamental, Estadística, Astronomía, Mecánica y Geodesia, Metodología y Didáctica de la Matemática, Ciencias de la Computación.

§4.- Importancia y papel del Algebra en la Matemática actual.

La palabra *álgebra* proviene del nombre de un tratado del matemático y astrónomo Mohammed al-Kharizmi, que vivió en el siglo IX. Su tratado sobre álgebra llevaba por título al-jebrw◠almugabala, que significa "transposición y eliminación". Por transposición se entiende la transferencia de términos al otro miembro de la ecuación, y por eliminación la cancelación de términos iguales en ambos miembros.

La palabra árabe al-jebr se convirtió en álgebra al transcribirla al latín, mientras que al-mugabala fué desechada, lo cual explica el término moderno de *álgebra* para esta disciplina.

El origen de este término responde muy bien al contenido real de la ciencia misma. El Algebra es en esencia la doctrina de las operaciones matemáticas consideradas formalmente desde un punto de vista general, con abstracción de los números concretos, y sus problemas están relacionados fundamentalmente con las reglas formales para la transformación de expresiones y la resolución de ecuaciones.

Más tarde, Omar Khayyan definió el Algebra como la ciencia de resolver ecuaciones. Esta definición no tuvo su significado hasta finales del siglo XIX, cuando el Algebra, junto con la teoría de ecuaciones, tomó nuevos derroteros, modificando esencialmente su carácter, pero no ese espiritu de generalidad que posee como ciencia de las operaciones formales.

El Algebra contemporánea es èl estudio de las operaciones, de las reglas de cálculo. Pero no se circunscribe como el Algebra clásica, al estudio de las propiedades de las operaciones con números, sino que aspira a investigar propiedades de operaciones con elementos de una naturaleza mucho más general. Esta tendencia viene dictada por necesidades de orden práctico. Por ejemplo, en Mecánica sumamos fuerzas, velocidades, rotaciones, etc. Si para un conjunto dado de objetos se definen ciertas operaciones que satisfacen ciertas propiedades, se dice entonces que se ha definido una estructura algebraica. El actual punto de vista sobre el Algebra consiste en considerarla como el estudio de las diferentes estructuras algebraicas. Puede considerarse que la noción de estructura aparece, con la definición por Cayley en 1854, del concepto de grupo abstracto y se desarrolla hasta la teoría de categorías actual, desarrollada en los últimos cuarenta años, que proporciona el marco correcto para el desarrollo de técnicas de gran importancia como la homología, que reunen aspectos aislados que habían ido apareciendo al profundizar en problemas de teoría de grupos, anillos, módulos, etc. El primer trabajo en el que se enfoca el Algebra desde el punto de vista de las estructuras es la famosa obra de Van der Waerden: "Modern Algebra", de importancia capital para el desarrollo algebraico posterior.

Hablemos ahora un poco de la influencia del Algebra en otras ramas de la Matemática, y en otras ciencias en general. El Algebra no es una ciencia aplicada en el sentido que tienen éstas hoy en día, sino una ciencia pura. Las ciencias aplicadas tienen, en su acepción usual, dos características que las definen, la de resolver problemas concretos del mundo que nos rodea y la de tomar prestado para este fin, un cuerpo de doctrina ya elaborado. El Algebra no depende de nada, salvo de la teoría

de conjuntos, de la que, en última instancia, depende la Matemática toda, y además es una ciencia pura porque tiene su propia problemática, independiente de los fenómenos de la vida real. Pero el Algebra si es una ciencia que se aplica. Ella presta a otras ramas de la Matemática y a otras ciencias en general, sus estructuras para lograr descripciones formales que las aclaren y potencien nuevos descubrimientos. Bien conocidas son, por ejemplo, las aplicaciones a la Física de la teoría de grupos y álgebras de Lie.

Y es que en el fondo de todo objeto matemático o colección de objetos, se encuentra la estructura algebraica. Por eso la casi totalidad de las ramas matemáticas usan de los teoremas del Algebra en su propio beneficio. Pero esta dependencia del Algebra no es como la dependencia, por ejemplo, de la Lógica. La Lógica suministra el esquema de razonamiento verdadero, pero ahí para su misión. El Algebra en cambio, como ciencia positiva que es, suministra a otras partes de la matemática, resultados positivos que ellas usan para sacar sus conclusiones, asimismo positivas.

§5.- Campos de investigación y desarrollo actual del Algebra. El Departamento de Algebra de la Complutense.

Un brevísimo esbozo histórico del desarrollo de la investigación matemática española nos mostraría, situándonos en los finales del pasado siglo, un panorama anquilosado y anclado en un pasado remoto. Sólo insólitos esfuerzos personales, como el de Reyes y Prosper, y la visión de futuro de Algunos prohombres de la Matemática, que empiezan a remover el terreno, hacen vislumbrar posibles cosechas futuras. Hombres como Echegaray. Eduardo Torroja, ó García de Galdeano conectan con la ciencia europea, comienzan a enseñar la matemática de su tiempo y no la de siglos pasados, fundan revistas matemáticas, crean la Real Sociedad Matemática Española, llegan a interesar y comunicar su entusiasmo a los estudiosos y envian a éstos al extranjero a formarse en las mismas fuentes. Fruto, y al mismo tiempo continuidad de estos esfuerzos, es Rey Pastor que, en su Laboratotio Matemático de la Junta de Ampliación de Estudios, consigue ya formar un equipo de investigadores.

Terminada la guerra, surgen unos cuantos jóvenes matemáticos que trabajan individualmente con notables aportaciones y que poco a poco empiezan a constituir a su alrededor grupos de investigación. El escaso número de matemáticos que salía de nuestras universidades, de las que únicamente las de Madrid, Barcelona y Zaragoza, y bastante después la de Santiago, impartían esos estudios, se ve luego altamente incrementado al crearse secciones y facultades de matemáticas en prácticamente todas las universidades. Esto favorece el número de los

que se dedican a la investigación y empieza ya a realizarse ésta de un modo continuado y en mayor escala.

Este proceso de formación de investigadores ha llevado en ocasiones a la constitución de auténticas escuelas: un primer maestro, dedicado a una determinada rama de su ciencia, la va desarrollando con sus discípulos, que a su vez prosiguen la tarea formadora de nuevos investigadores incipientes.

Por otra parte, la comunidad matemática española aparece fuertemente relacionada gracias a la celebración, desde hace treinta años, de una Reunión anual que va turnando su sede por las distintas Universidades. También se han vinculado a estos congresos los matemáticos portugueses y así se intercala una reunión en una universidad lusitana por cada dos reuniones en universidades españolas. Estas reuniones, aparte de favorecer el intercambio y la colaboración entre investigadores y docentes de la matemática, son la mejor plataforma para comenzar a darse a conocer nuestros jóvenes valores, cuya presencia cada vez más activa en ellas permite predecir un interesante futuro. También cada cuatro años se celebra el Congreso de la Agrupación de Matemáticos de Expresión Latina.

Fruto de estos congresos de distintos tipos es la publicación de las actas correspondientes, que recogen las comunicaciones presentadas. Otras publicaciones son dignas de ser anotadas. Muchos departamentos universitarios publican de modo informal sus trabajos, aunque algunos ya en forma seriada, como la Sección de Matemáticas de la Universidad Autónoma de Barcelona, ó el departamento de Geometría y Topología, ó la revista "Alxebra", ambos en Santiago, ó la colección de Monografías y Memorias de Matemáticas del Instituto Jorge Juan por citar algunos. Asimismo, la revista de las Academias de Madrid, Barcelona y Zaragoza dedican parte de sus números a artículos de investigación matemática. Finalmente, como revistas matemáticas de carácter general, hay que destacar "Collectanea Mathemática", editada por la Universidad de Barcelona y "Revista Matemática Iberoamericana", publicada por la Real Sociedad Matemática Española en colaboración con el C.S.I.C.

Observamos con orgullo que la producción matemática española aparece como muy superior si se la relaciona con la inversión a ella destinada, encontrándose potencialmente en situación de despegue, en cuanto que esta inversión no necesita ser muy cuantiosa. Fundamentalmente, dicha inversión estaría destinada a promover estancias de investigadores españoles en centros extranjeros que sirvieran para catalizar la labor de nuestros investigadores.

Informaremos ahora brevemente sobre las investigaciones de Algebra en España. El desarrollo de la investigación en Algebra se centra en torno a cuatro grandes polos, que pueden ser recogidos con las denominaciones siguientes: 1) geometría algebraica, 2) teoría de grupos, 3) teoría de categorías, 4) teoría de números. Estos bloques de trabajo han nacido de cuatro grupos o escuelas que se localizan, respectivamente, en Madrid, Zaragoza, Santiago y Barcelona, dando lugar a nuevos núcleos al desplazarse alhunos de sus miembros a otros centros del país.

La escuela de geometría algebraica radicada en Madrid, en la Universidad Complutense y en el Instituto Jorge Juan del C.S.I.C., ha producido nuevos grupos localizados en las Universidades de Valladolid, Sevilla, Santander, La Laguna, Málaga, y en parte en Zaragoza. Independientemente hay grupos de trabajo en las Universidades de Barcelona, Salamanca y Badajoz, que tocan los mismos temas. La teoría de grupos, cultivada en la Universidad de Zaragoza, se ha extendido después, a través de sus miembros, a las Universidades de Santander y Valencia. La teoría de categorías y álgebra homológica de Santiago, tiene, igualmente, su ramificación en las Universidades de Granada, Málaga y Murcia. Y los grupos de trabajo sobre teoría de números se encuentran en Barcelona, Bilbao y Madrid. Lo anterior es una simplificación que no excluye el hecho de que también se trabaje algo sobre teoría de grupos en Valladolid o sobre teoría de categorías en Zaragoza.

Como hemos dicho anteriormente, en el Departamento de Algebra de la Universidad Complutense de Madrid existen dos grandes temas de investigación: La teoría de números y la geometría algebraica, dividida esta última en dos vertientes: la geometría algebraica real y la teoría de la resolución de singularidades.

Observaremos como conclusión final que el cultivo del álgebra es sus distintas ramas es ya un hecho consolidado entre nuestros estudiosos y los grupos jóvenes que se han incorporado a él hacen prometer el desarrollo de esta disciplina.

Rev. Academia de Ciencias. Zaragoza. 49 (1994)

ERROR BOUND REPRESENTATIONS OF CHEBYSHEV-HALLEY TYPE METHODS IN BANACH SPACES.

Dong Chen [1], I.K. Argyros [2] and Q.S. Qian [3]

(1) Department of Mathematical Sciences, Univ. of Arkansas, Fayetteville, Arkansas 72701, USA.

(2) Department of Mathematics, Cameron University, Lawton, Oklahoma 73505, USA.

(3) Department of Mathematics, Univ. of Kentucky, Lenxington, Kentucky 40506, USA.

Abstract

We provide a Ostrowski-Kantorovich convergence theorem under standard Ostrowski-Kantorovich conditions for a family of Chebyshev-Halley type methods in Banach spaces. We propose the upper error bound and lower bound for this family with a real parameter λ ($0 \leq \lambda < 2$). We also discuss sufficient asymptotic error bounds for the methods.

1. Introduction.

Ostrowski-Kantorovich convergence of the Chebyshev- Halley iterative methods in Banach space setting was studied by M.A. Mertvecova [13] , M. Altman [1,2] , B. Doring [8] and R.A. Safiev [18,19]. Later, T. Yamamoto [21,22] developed a theory of a cubic optimal operator and applied this theory to the study of Ostrowski-Kantorovich convergence for the Halley method. In recent years V. Candela and A. Marquina [6] provided Ostrowski-Kantorovich type convergence theorems for both Chebyshev-Halley methods in Banach spaces, by employing a special technique which is called the recurence relation. They also convinced that both methods are applicable by providing many numerical examples. This year, S. Kanno [11] examined two such convergence theorems for Halley method by Safiev and Yamamoto [21].

He points out that Yamamoto's assumptions are weaker than that of Safiev and the error bound is finer than that of Safiev. In this paper,

we consider similar problems for a family of Chebyshev-Halley type methods which contain both Chebyshev and Halley methods as specific cases. By employing classical analysis techniques and under similar assumptions of the Newton-Ostrowski-Kantorovich theorem [9,10,12,14, 15,16,17,20,21] , we give the sufficient conditions and a complete representation of error bound based on the initial information of the nonlinear operator equation $P(X) = 0$ for Chebyshev-Halley type methods. It means that we can provide the convergence and error bound for Chebyshev and Halley methods based on the quadratic optimal operator. But we also point out that there is a method for which the quadratic optimal operator does not work. Only Yamamoto's third order optimal operator can be applied in order to present the convergence and error bound. Finally, we discuss some sufficient asymptotic error bounds for the methods for all parameters λ in [0,2], and provide numerical examples.

2. The extension of Chebyshev-Halley type methods in Banach spaces.

Consider a nonlinear operator equation of the form

$$P(X) = 0 \tag{2.1}$$

where $P: D_0 \subset X_B \to Y_B$ is a nonlinear mapping defined on an open convex domain D_0 of a Banach space X_B with values in a Banach space Y_B. Under certain conditions and based on the original Chebyshev-Halley type methods [23], we can define an equivalent form for the family in Banach space for all $n \geq 0$:

$$\begin{aligned} Y_n &= X_n - P'(X_n)^{-1} P(X_n) \\ H(X_n, Y_n) &= P'(X_n)^{-1} P''(X_n)(Y_n - X_n) \\ X_{n+1} &= Y_n - \frac{1}{2}\left[I + \frac{\lambda}{2} H(X_n, Y_n)\right] H(X_n, Y_n)(Y_n - X_n), \end{aligned} \tag{2.2}$$

where λ is a real parameter. In section 4, we will prove that the sequence $\{X_n\}_{n=1}^{\infty}$ generated by (2.2) is well-defined and convergent to the solution of the nonlinear operator equation (2.1) under standard Ostrowski-Kantorovich assumptions.

For convenience and brevity, in the following section 3, we introduce the functions

$$Z_n = X_n + t(Y_n - X_n) \qquad Q_n = \left[I + \frac{\lambda}{2} H(X_n, Y_n)\right] \tag{2.3}$$

3. The Ostrowski-Kantorovich theorem.

First we need a lemma as a useful tool for estimation.

Lemma 3.1. Let P be a nonlinear operator on an open convex domain D_0 of a Banach space X_B to another Banach space. Suppose that P has 2nd order continuos Frechet derivatives on D_0. Then the $P(X_{n+1})$ together with the sequence $\{X_n\}_{n=0}^{\infty}$ generated by (2.2) have the following identity for all $n \geq 0$ and $0 \leq l \leq 2$:

$$P(X_{n+1}) = \int_0^1 P''[Y_n + t(X_{n+1} - Y_n)](1-t)dt(X_{n+1} - Y_n)^2 -$$

$$\frac{1}{2}\int_0^1 P''(Z_n)[1-\lambda(1-t)]dt(Y_n - X_n)\, Q_n\, H(X_n, Y_n)(Y_n - X_n) +$$

$$\int_0^1 \left\{ P''(Z_n)(1-t) - \frac{1}{2} P''(X_n) \right\} dt(Y_n - X_n)\, Q_n\, (Y_n - X_n) \tag{3.2}$$

Now we can state our main results:

Theorem 3.3 Let $P : D_0 \quad X_B \to Y_B$, X_B, Y_B are real or complex Banach spaces, and D_0 is an open convex domain. Assume that P has 2nd order continuous Frechet derivatives on D_0 and satisfies the following standard Ostrowski-Kantorovich conditions:

$$\| P''(X) \| \leq M \ , \ \| P''(X) - P''(Y) \| \leq N \| X - Y \| \ , \text{for all } X, Y \in D_0 \tag{3.4}$$

For a given initial value $X_0 \in D_0$, assume that $P(X_0)^{-1}$ exists and

$$\| P'(X_0)^{-1} \| \leq \beta \ , \quad \| Y_0 - X_0 \| \leq \eta \ , \tag{3.5}$$

$$M\left[1 + \frac{2N}{3(2-\lambda)M^2\beta}\right]^{1/2} \leq K \ , \ 0 \leq \lambda < 2 \ , \tag{3.6}$$

$$h = K\beta\eta \leq \begin{cases} 0.485 \\ 0.5 \end{cases} \quad \text{if} \quad \begin{cases} 0 \leq \lambda \leq 1 \\ 0 \leq \lambda \leq 2 \end{cases} \ , \tag{3.7}$$

$$\overline{S(X_0, t^*)} \subset D_0 , \tag{3.8}$$

where $\overline{S(x, r)} = \{ x' \in X \mid \| x' - x \| \leq r \}$, and set

$$g(t) = \tfrac{1}{2}Kt^2 - \tfrac{1}{\beta}t + \tfrac{\eta}{\beta} , \quad g(t) = 0 , \tag{3.9}$$

$$t^* = \frac{1 - \sqrt{1 - 2h}}{h}\eta , \tag{3.10}$$

$$\theta = (1 - \sqrt{1 - 2h})/(1 + \sqrt{1 - 2h}) , \tag{3.11}$$

where t^* is the smallest root of the equation (3.9). Then the Chebyshev -Halley type procedures (2.2) are convergent for all $0 \leq \lambda < 2$. Also X_n, $Y_n \in \overline{S(X_0, t^*)}$, for all $n \geq 0$. The limit X^* is a solution of the equation (2.1). We also have the following error estimates and the optimal error bounds:

$$\| X_n - X^* \| \leq t^* - t_n , \text{ for all } n \geq 0 , \tag{3.12}$$

$$\| Y_n - X^* \| \leq t^* - s_n , \text{ for all } n \geq 0 , \tag{3.13}$$

Now, putting

$$\nu = 3^n - 1 , \quad \Gamma = \sqrt{2 - \lambda} , \quad \Delta = \sqrt{\frac{2 + \theta}{1 + 2\theta}} ,$$

we obtain

$$\frac{(1-\theta^2)\eta(\Gamma\theta)^\nu}{1-\frac{(\Gamma\theta)^{\nu+1}}{\Gamma}} \leq t^* - t_n \leq \frac{(1-\theta^2)\eta\theta^\nu}{1-\theta^{\nu+1}} \tag{3.14}$$

for all λ in the interval $[1,2)$, and

$$\frac{(1-\theta^2)\eta\theta^\nu}{1-\theta^{\nu+1}} \leq \frac{(1-\theta^2)\eta(\Delta\theta)^\nu}{1-\frac{(\Delta\theta)^{\nu+1}}{\Delta}} \leq t^* - t_n \leq \frac{(1-\theta^2)\eta(\theta\sqrt{2})^\nu}{1-\frac{(\theta\sqrt{2})^{\nu+1}}{\sqrt{2}}} \tag{3.15}$$

for all λ in the interval $[0,1]$, where $\{t_n\}_{n=0}^{\infty}$ and $\{s_n\}_{n=0}^{\infty}$ are defined as

$$s_n = t_n - \frac{g(t_n)}{g'(t_n)} \tag{3.16}$$

$$h_g(t_n , s_n) = \frac{g''(t_n)(s_n - t_n)}{g'(t_n)}$$

$$t_{n+1} = s_n - \frac{h_g(t_n , s_n)(s_n - t_n)}{2 + \lambda h_g(t_n , s_n)} \qquad (3.16)$$

Proof. It suffices to show that the following item is true for all n by mathematical induction:

(I_n) $\quad X_n \in \overline{S(X_0 , t_n)}$;

(II_n) $\quad \| P'(X_n)^{-1} \| \le -g'(t_n)^{-1}$;

(III_n) $\quad \| Y_n - X_n \| \le s_n - t_n$;

(IV_n) $\quad Y_n \in \overline{S(X_0 , s_n)}$;

(V_n) $\quad \| X_{n+1} - Y_n \| \le t_{n+1} - s_n$

Proof. It is easy to check in the case of $n = 0$ by initial conditions. Now assume that (I_n) - (V_n) are true for a fixed $n \ge 1$. Then

(I_{n+1}): $\| X_{n+1} - X_0 \| \le \| X_{n+1} - Y_n \| + \| Y_n - X_n \| + \| X_n - X_0 \| \le$

$$\le (t_{n+1} - s_n) + (s_n - t_n) + (t_n - t_0) = t_{n+1} .$$

$$(II_{n+1}) : P'(X_{n+1}) - P'(X_0) = \int_0^1 P''[X_0 + t(X_{n+1} - X_0)]dt(X_{n+1} - X_0) ,$$

so

$$\| P'(X_{n+1}) - P'(X_0) \| \le M \| X_{n+1} - X_0 \| \le K(t_{n+1} - t_0) = Kt_{n+1} < Kt^* =$$

$$= K\eta \frac{1 - \sqrt{1 - 2h}}{h} = \frac{1 - \sqrt{1 - 2h}}{\beta} \le \frac{1}{\beta} \le \frac{1}{\| P'(X_0)^{-1} \|} ,$$

and by Banach Theorem, $P'(X_{n+1})^{-1}$ exists and

$$\| P'(X_{n+1})^{-1} \| \le \frac{\| P(X_0)^{-1} \|}{1 - \| P'(X_0)^{-1} \| \| P'(X_{n+1}) - P'(X_0) \|}$$

$$\le \frac{\beta}{1 - \beta K \| X_{n+1} - X_0 \|} \le \frac{1}{(1/\beta) - K(t_{n+1} - t_0)} = \frac{1}{(1/\beta) - Kt_{n+1}} = - g'(t_{n+1})^{-1}$$

(III_{n+1}): Putting $\Phi_n = \| X_n - X_0 \|$, $\Psi_n = \| Y_n - X_n \|$, and by using the identity (3.2), we can estimate $P(X_{n+1})$ to obtain:

$$\| P(X_{n+1}) \| \le \frac{M}{2} \| X_{n+1}-Y_n \|^2 + \left[\frac{1}{2} - \frac{\lambda}{4} \right] \frac{\dfrac{M^2\Psi^3}{(1/\beta) - M\Phi}}{1 - \dfrac{\lambda}{2} \dfrac{M\Psi}{(1/\beta) - M\Phi}}$$

$$+ \frac{N}{6} \frac{\Psi^3}{1 - \dfrac{\lambda}{2} \dfrac{M\Psi}{(1/\beta) - M\Phi}} \le \frac{M}{2} \| X_{n+1}-Y_n \|^2 + \frac{\dfrac{\left[\dfrac{(2-\lambda)M^2}{4} + \dfrac{N}{6\beta} \right] \Psi^3}{(1/\beta) - M\Phi}}{1 - \dfrac{\lambda}{2} \dfrac{M\Psi}{(1/\beta) - M\Phi}}$$

$$\le \frac{K}{2}(t_{n+1}-s_n)^2 + \frac{(2-\lambda)K^2}{4} \frac{\dfrac{(s_n-t_n)^3}{(1/\beta)-Kt_n}}{1 - \dfrac{\lambda}{2} \dfrac{K(s_n-t_n)}{(1/\beta)-Kt_n}} = g(t_{n+1})$$

and so

$$\| Y_{n+1}-X_{n+1} \| \le \| P'(X_{n+1})^{-1} \| \| P(X_{n+1}) \| \le -\frac{g(t_{n+1})}{g'(t_{n+1})} = s_{n+1} - t_{n+1}$$

(IV_{n+1}): $\| Y_{n+1}-X_0 \| \le \| Y_{n+1}-X_{n+1} \| + \| X_{n+1}-Y_n \| + \Psi + \Phi$

$$\le (s_{n+1}-t_{n+1}) + (t_{n+1}-s_n) + (s_n-t_n) + (t_n-t_0) = s_{n+1}$$

(V_{n+1}): $\| \frac{\lambda}{2} P'(X_{n+1})^{-1} P''(X_{n+1})(Y_{n+1}-X_{n+1}) \| \le \frac{\lambda}{2} \| P'(X_{n+1})^{-1} \| \| P''(X_{n+1}) \| *$

$$* \| Y_{n+1}-X_{n+1} \| \le \frac{\lambda}{2} \frac{K(s_{n+1}-t_{n+1})}{-g'(t_{n+1})} \le \frac{K(s_{n+1}-t_{n+1})}{(1\backslash\beta) - Kt_{n+1}} < 1$$

thus $\left[I + \frac{\lambda}{2} P'(X_{n+1})^{-1} P''(X_{n+1})(Y_{n+1}-X_{n+1}) \right]$ exists, and

$$\left\| \left[I + \frac{\lambda}{2} P'(X_{n+1})^{-1} P''(X_{n+1})(Y_{n+1}-X_{n+1}) \right]^{-1} \right\| \le$$

$$\le \left[1 - \frac{\lambda}{2} \| P'(X_{n+1})^{-1} \| \| P''(X_{n+1}) \| \| Y_{n+1}-X_{n+1} \| \right]$$

$$\le \left[1 + \frac{\lambda}{2} g'(t_{n+1})^{-1} K(s_{n+1}-t_{n+1}) \right]^{-1} \le \left[1 + \frac{\lambda}{2} g'(t_{n+1})^{-1} g''(t_{n+1}) * \right.$$

$$\left. * K(s_{n+1}-t_{n+1}) \right]^{-1} = \left[1 + \frac{\lambda}{2} h_g(t_{n+1}, s_{n+1}) \right]$$

From (2.2) we have

$$X_{n+2}-Y_{n+1} = -\frac{1}{2}\left[I+\frac{\lambda}{2}H(X_{n+1}, Y_{n+1})\right]^{-1} H(X_n, Y_n)(Y_n-X_n)$$

and then

$$\|X_{n+2}-Y_{n+1}\| \le -\frac{1}{2}\left\|\left[I+\frac{\lambda}{2}H(X_{n+1}, Y_{n+1})\right]^{-1}\right\| \| P'(X_{n+1})^{-1} \| *$$

$$\| P''(X_{n+1}) \| \| Y_{n+1}-X_{n+1} \|^2 \le -\frac{1}{2}\left[I+\frac{\lambda}{2}h_g(t_{n+1}, s_{n+1})\right]^{-1} h_g(t_{n+1}, s_{n+1}) *$$

$$* (s_{n+1}-t_{n+1}) = t_{n+2} - s_{n+1}$$

Now we are ready to prove (3.14) and (3.15). Notice that

$$g(t_n) = \frac{K}{2}(t^* - t_n)(t^{**} - t_n),$$

$$g'(t_n) = -\frac{K}{2}\left[(t^* - t_n) + (t^{**} - t_n)\right].$$

Denote $a_n = t^* - t_n$, $b_n = t^{**} - t_n$, Then we have

$$g(t_n) = \frac{K}{2} a_n b_n, \quad g'(t_n) = -\frac{K}{2}(a_n + b_n), \quad b_n = a_n + (1-\theta^2)\frac{\eta}{\theta}.$$

Now by (3.16), we have

$$a_n = a_{n-1} - \frac{a_{n-1}b_{n-1}(a_{n-1} + b_{n-1})^2 + (1-\lambda)a_{n-1}^2 b_{n-1}^2}{(a_{n-1} + b_{n-1})^3 - \lambda a_{n-1}b_{n-1}(a_{n-1} + b_{n-1})}$$

$$= \frac{a_{n-1}^4 + (2-\lambda)a_{n-1}^3 b_{n-1}}{(a_{n-1} + b_{n-1})^3 - \lambda a_{n-1}b_{n-1}(a_{n-1} + b_{n-1})}$$

By a similar way we should have an expression of b_n:

$$b_n = \frac{b_{n-1}^4 + (2-\lambda)b_{n-1}^3 a_{n-1}}{(a_{n-1} + b_{n-1})^3 - \lambda a_{n-1}b_{n-1}(a_{n-1} + b_{n-1})}$$

So, we obtain

$$\frac{a_n}{b_n} = \{a_{n-1}/b_{n-1}\}^3 \frac{a_{n-1} + (2-\lambda)b_{n-1}}{b_{n-1} + (2-\lambda)a_{n-1}} = \{a_{n-1}/b_{n-1}\}^3 \frac{\frac{a_{n-1}}{b_{n-1}} + (2-\lambda)}{1 + (2-\lambda)\frac{a_{n-1}}{b_{n-1}}}$$

Case (i) : $0 \le \lambda \le 1$. Note that $0 \le \frac{a_{n-1}}{b_{n-1}} \le \theta \le 1$, so

$$1 \le \Delta^2 \le \frac{\frac{a_{n-1}}{b_{n-1}} + (2-\lambda)}{1 + (2-\lambda)\frac{a_{n-1}}{b_{n-1}}} \le 2 .$$

That is

$$\{a_{n-1}/b_{n-1}\}^3 \le \Delta^2\{a_{n-1}/b_{n-1}\}^3 \le \{a_n/b_n\} \le 2\{a_{n-1}/b_{n-1}\}^3$$

Then we solve this equation for a_n by using the fact that

$b_n = a_n + (1-\theta^2)\frac{\eta}{\theta}$.

By (3.15), it is easy to see that

$$\frac{(1-\theta^2)\eta\theta^\nu}{1 - \theta^{\nu+1}} \le \frac{(1-\theta^2)\eta(\Delta\theta)^\nu}{1 - \frac{(\Delta\theta)^{\nu+1}}{\Delta}} \le a_n = t^* - t_n \le \frac{(1-\theta^2)\eta(\theta\sqrt{2})^\nu}{1 - \frac{(\theta\sqrt{2})^{\nu+1}}{\sqrt{2}}} .$$

Case(ii) : $1 \le \lambda < 2$. By a similar method as above and taking into account (3.14), we have the following error bounds:

$$\frac{(1-\theta^2)\eta(\Gamma\theta)^\nu}{1 - \frac{(\Gamma\theta)^{\nu+1}}{\Gamma}} \le a_n = t^* - t_n \le \frac{(1-\theta^2)\eta\theta^\nu}{1 - \theta^{\nu+1}}$$

4. Some sufficient asymptotic error bounds.

We discuss some characterizations of the methods (2.2) under the quadratic optimal operator. We observe that if replace P(X) by g(t) in (3.2), then we should have

$$g(t_{n+1}) = \frac{K}{2}(t_{n+1} - s_n)^2 + \frac{\left[\frac{\lambda}{4} - \frac{1}{2}\right]K^2(s_n - t_n)^3}{g'(t_n)\left[1 + \frac{\lambda}{2}h_g(t_n , s_n)\right]} \qquad (4.1)$$

By using the fact that

$$t_{n+1} - s_n = -\frac{1}{2}\frac{\frac{g''(t_n)}{g'(t_n)}(s_n-t_n)^2}{1 + \frac{\lambda}{2} h_g(t_n , s_n)}$$

and putting

$$\Omega = 1 + \frac{\lambda K}{2g'(t_n)} (s_n-t_n)$$

we now get

$$g(t_{n+1}) = \frac{K^2(s_n-t_n)^3}{\Omega g'(t_n)} \left[\frac{K(s_n-t_n)}{8\Omega g'(t_n)} + \frac{\lambda}{4} - \frac{1}{2}\right] \tag{4.2}$$

(i) If $\lambda = 2$, then

$$g(t_{n+1}) = \frac{K^3(s_n-t_n)^4}{8\Omega g'(t_n)^2}$$

and the sufficient asymptotic error bound for $n \to \infty$, is given by

$$\lim \frac{g(t_{n+1})}{(s_n-t_n)^4} = \frac{K^3\beta^2}{8(1-2h)} \tag{4.3}$$

(ii) If $0 \le \lambda < 2$, then the sufficient asymptotic error bound for $n \to \infty$, is also given by

$$\lim \frac{g(t_{n+1})}{(s_n-t_n)^3} = \frac{(2-\lambda)K^2\beta}{4\sqrt{1-2h}} \tag{4.4}$$

5. Applications.

In this section, first we use the theorem 3.3 to suggest some new approaches to the solution of quadratic integral equations of the form:

$$X(s) = Y(s) + \alpha X(s) \int_0^1 q(s,t)\, X(t)dt \tag{5.1}$$

in the space $X_B = C[0 , 1]$, of all continuous functions on the interval $[0 , 1]$, with norm for $0 \le s \le 1$

$$\| X \| = \max_{0 \le s \le 1} | X(s) | \tag{5.2}$$

Here we assume that α is a real number called the "albedo" for scattering and the kernel $q(s,t)$ is a continuous function of two variables, with $0 \leq s , t \leq 1$, and satisfying

(i) $0 < q(s,t) < 1 , \quad 0 \leq s , t \leq 1 ;$

(ii) $q(s,t) + q(t,s) = 1 , \quad 0 \leq s , t \leq 1 ;$

The function $Y(s)$ is given continuous function defined on the interval $[0,1]$, and $X(s)$ is the unknown function sought in $[0,1]$. Equations of this type are related with the work of s. Chandrasekhar [7], and arise in the theories of radiative transfer, neutron transport and in the kinetic theory of gasses. These exists an extensive literature on equations like (5.1) under various assumptions on the kernel $q(s,t)$ and α is a real or complex number. One can refer to the recent work in [3,4,5] and thereferences there. Here we demonstrate that the theorem via the iterative procedures (2.2) provide existence results for (5.1). Moreover the iterative procedures (2.2) converge faster to the solution that all the previous known ones. Furthermore a better information on the location of the solution is given. Note that cost is not higher that the corresponding one of previous methods. For simplicity, we shall assume that

$$q(s,t) = \frac{s}{s + t} , \quad \text{for all} \quad 0 \leq s , t \leq 1 .$$

Notice that $q(s,t)$ satisfies (i) and (ii) above. Let us now choose $Y(s) = 1$ for all s in $[0,1]$ and define the operator P on $X_B = C[0,1]$ by

$$P(X) = \alpha X(s) \int_0^1 \frac{s}{s + t} X(t)\, dt - X(s) + 1 \qquad (5.3)$$

Note that every root of the equation $P(X) = 0$ satisfies the equation (5.3) Set $X_0(s) = 1$ and $\alpha = 0.25$, use the definition of the first and second Frechet derivatives of the operator P to obtain

$$M\, 2|\alpha|\, \max \left| \int_0^1 \frac{s}{s + t} dt \right| = (2 \ln 2)|\alpha| = 0.34657359 ,$$

$N = 0$

$K = M$

$\beta = \| P'(1)^{-1} \| = 1.53039421$

$\eta \geq \| P'(1)^{-1} P(1) \| \geq 0.26519711$

$$h = K\beta\eta = 0.14065901 < \frac{1}{2}$$

$$t^* = 0.28704852$$

$$\theta = 0.o8239685$$

$$\| X_n(\lambda) - X^* \| \leq \frac{(1-\theta^2)\eta\theta^v}{1-\theta^{v+1}} = \frac{0.26339662[0.08239685]^v}{1 - [0.08239685]^{v+1}}$$

for $1 \leq \lambda < 2$ and

$$\| X_n(\lambda) - X^* \| \leq \frac{(1-\theta^2)\eta(\theta\sqrt{2})^v}{1 - \frac{(\theta\sqrt{2})^{v+1}}{\sqrt{2}}} .= \frac{0.26339662\,[0.116526742]^v}{1 - \frac{1}{\sqrt{2}}[0.116526742]^{v+1}}$$

for all $0 \leq \lambda \leq 1$ and all $n \geq 0$, which shows that X^* is unique in $S(X_0,t^*)$. We now discuss the determination of the parameter λ so that the iterative procedures (2.2) will produce better solutions by spending thesame amount of computations. Our numerical example do convince the above theoretical conclusions. Let us consider $P(X) = X^3 - 2X - 5$, where $X^* = 2.094551481$, and

$$E_1(\lambda) = \| X_1(\lambda) - X^* \| .$$

Then we have the following numerical results:

Table 5.4

λ	X_0	X_1	$E_0(\lambda)$	$E_1(\lambda)$
0.0	2.0	2.094	0.95×10^{-1}	0.55×10^{-3}
1.0	2.0	2.0943396	0.95×10^{-1}	0.21×10^{-3}
2.0	2.0	2.0946429	0.95×10^{-1}	0.91×10^{-4}
3.0	2.0	2.0949152	0.95×10^{-1}	0.36×10^{-3}
4.0	2.0	2.0951612	0.95×10^{-1}	0.61×10^{-3}

References

[1] M. Altman: *Iterative Methods of Higher Order,* Bull. Acad. Polon. Sci. Math. Ast. Phys., 9(1961), 63-68.

[2] M. Altman: *Conerning the Method of Tangent Hyperbolas for Operator Equations,* Bull. Acad. Polon. Sci. Ser. Math. Astr. Phys. 9(1961), 633-637.

[3] I.K. Argyros: *Quadratic Equations and Applications to Chandrasekar' 's and Related Equations,* Bull. Austral. Math. Soc., 32(1988), 275 - 292.

[4] I.K. Argyros: *On a Class of Nonlinear Inteqral Equations Arising in Neutron Transport,* Aequations Mathematicae, 36(1988), 99-111.

[5] I.K. Argyros: *The Secant Method in Generalized Banach, Spaces,* Applied Math. & Comp., 39(1990). 111-121.

[6] V. Candela and A. Marquina: *Recurrence Relations for Rational Cubic Methods I: The Halley Method,* Computing, 44(1990), 169-184.

[7] S. Chandrasekhar: *Radiative Transfer,* Dover Pub., New York, 1960.

[8] B. Doring: *Einige Staze Uber das Verfahren der tangierenden Hyper-beln in Banach-Raumen,* Aplikace Math., 15(1970), 418-464.

[9] W.B. Gragg and R.A. Tapia: *Optimal Error Bounds for Newton-Kantorovich Theorem,* SIAM J. Numer. Anal., 11(1974), 10-13.

[10] V.S. Grebenjuk: *Application of Majorant Method to a Class of Iterative Methods,* Ukraninski Mat. Z., 18(1966), 102-106.

[11] S. Kanno: *Convergence Theorems for the Method of Tangent Hyper- bolas,* Mathematica Japonica, 37: 4(1992), 711-722.

[12] L.V. Kantorovich and G.P. Akilov: *Functional Analysis in Normed Spaces,* Pergamon Press, New York, 1964.

[13] M.A. Mertvecova: *An Analog of the Process of Tangent Hyperbolas for General Functional Equations,* Dokl. Akad. Naut. SSSR, 88(1953), 622-614.

[14] J.M. Ortega and W.C. Rheinboldt: *Iterative Solution of Nonlinear Equations in Several Variables,* Academic Press, New York, 1970.

[15] A.M. Ostrowski: *Solution of Equations in Euclidean and Banach Spaces,* Academic Press, New York, 3rd., 1973.

[16] F.A. Potra and V. Ptak: *Sharp Error Bounds for Newton's Process,* Numer. Math., 34(1980), 63-72.

[17] L.B. Rall: *Computational Solution of Nonlinear Operator Equations,* John Wiley & Sons, Inc., New York, 1969.

[18] R.A. Safiev: *The Method of Tangent Hyperbolas,* Sov. Math. Dokl. 4(1963), 482-485.

[19] R.A. Safiev: *On some Iterative Processes,* Z. Vyccisl. Mat. Fiz., 4(1964), 139-143.

[20] T. Yamamoto: *A Unified Derivation of Several Error Bounds for Newton's Process,* J. Compt. Appl. Math., 21(1985), 179-191.

[21] T. Yamamoto: *A Method for Finding Sharp Error Bounds for Newton's Methods under the Kantorovich Assumptions,* Numer. Math., 49(1986), 203-220.

[22] T. Yamamoto: *On the Method of Tangent Hyperbolas in Banach Spaces,* J. Comput. Appl. Math., 21(1988), 75-86.

[23] W. Werner: *Some Improvements of Classical Iterative Methods for the Solutions of Nonlinear Equations,* Lecture Notes in Math. Numerical Solution of Nonlinear Equations, Proceedings, Bremen, 878(1980), 427-440.

Rev. Academia de Ciencias. Zaragoza. 49 (1994)

Fourier exponential series for Fox's H-function of several variables.

S. D. BAJPAI

Department of Mathematics
University of Bahrain
P.O.Box 32038 Isa Town, BAHRAIN.
and
Institute for Basic Research
P.O.Box 1577, Palm Harbor, FL 34682, USA.

Abstract

In this paper, we establish one Fourier exponential series for Fox's H-function of several variables.

1. Introduction

In last three decades many mathematicians tried to present various Fourier series and expansions for the G and H-functions of two and several variables [6,7,8]. A serious study of their work reveals that almost all these Fourier series and expansions are not the Fourier series and expansions for the G and H-functions of two and more variables, but have been presented in a misleading form to appear as Fourier series and expansions for the G and H-functions of two or several variables and may be viewed as the manipulative forms of already known work on Meijer's G-function and Fox's H-function [6,7,8].

It is important to note that the Fourier series and expansions presented by these mathematicians [6,7,8] involve only one variable x and have been presented in terms of a single series, therefore these are Fourier series and expansions for a function of one variable. Any Fourier series or expansion for a function of several variables should involve several variables and should be presented in terms of a multiple series as discussed by Carslaw and Jaeger [2, pp.180-183] for Fourier series of two variables.

The object of this paper is to establish one Fourier exponential series for Fox's H-function of several variables with the help of a multiple integral evaluated in this paper.

Several mathematicians tried for generalizations of Fox's H-function [4] and defined Fox's H-function of two and several variables [7, pp.22-35 & 8, pp.82-98, 251-254].

In this paper, Fox's H-function of several variables [8, pp.251-254] will be represented as follows:

$$(1.1)\quad H\begin{bmatrix} z_1 \\ .. \\ z_r \end{bmatrix} \equiv H\begin{matrix} 0,n;m_1,n_1;\ ---;\ m_r,n_r \\ p,q;\ p_1,q_1;\ ---;\ p_r,q_r \end{matrix}\left[\begin{matrix} z_1 \\ .. \\ z_r \end{matrix}\middle|\begin{matrix} A;C_{p_1}\ ;\ ---;\ C_{p_r} \\ \\ B;D_{q_1}\ ;\ ---;\ D_{q_r} \end{matrix}\right] =$$

$$H\begin{matrix} 0,n;\ m_1,n_1;\ ---;\ m_r,n_r \\ p,q\ ;\ p_1,q_1;\ ---;\ p_r,q_r \end{matrix}\left[\begin{matrix} z_1 \\ .. \\ z_r \end{matrix}\middle|\begin{matrix} (a_j;\alpha_j^{(1)},---,a_j\alpha_j^{(r)})_{1,p_r} : (c_j^1,\gamma_j^1)_{1,p_1};--- \ (c_j^{(r)},\gamma_j^{(r)})_{1,p_r} \\ \\ (b_j;\beta_j^{(1)},---,b_j\beta_j^{(r)})_{1,q_r} : (d_j^1,\delta_j^1)_{1,q_1};---(d_j^{(r)},\delta_j^{(r)})_{1,q_r} \end{matrix}\right]$$

2. The multiple integral.

Putting

$$\omega = w + v \qquad \mu = w + v - 2 \qquad v = i(w - v)x$$

$$\omega(k) = w_k + v_k \qquad \mu(k) = w_k + v_k - 2 \qquad v(k) = i(w_k - v_k)x_k$$

the multiple integral to be evaluated is

$$(2.1)\quad \int_{-\pi/2}^{\pi/2} --- \int_{-\pi/2}^{\pi/2} (\cos x_1)^{\mu(1)} --- (\cos x_r)^{\mu(r)}\, e^{v(1)} --- e^{v(r)}$$

$$H\begin{bmatrix} z_1(e^{ix}\cos x_1)^{t_1} \\ \\ z_r(e^{ix}\cos x_r)^{t_r} \end{bmatrix} dx_1 --- dx_r = \frac{(\pi)^r}{w_1 + --- + w_r + v_1 + --- + v_r - 2r\,\Gamma(v_1) --- \Gamma(v_r)}$$

$$H\begin{matrix} 0,n;m_1,n_1+1\ ;---;\ m_r,n_r+1 \\ p,q;p_1+1,q_1+1;---;p_r+1,q_r+1 \end{matrix}\left[\begin{matrix} 2^{-t_1}z_1 \\ 2^{-t_r}z_r \end{matrix}\middle|\begin{matrix} A\ ;\ (-\mu(1),t_1),\ C_p\ ;---;\ (-\mu_r,t_r),\ C_{p_r} \\ B\ ;\ D_{q_1},(1-w_1,t_1)\ ;---\ ;\ D_{q_r},(1-w_r,t_1) \end{matrix}\right]$$

$$\mathrm{Re}(w_i + v_i) + t_i \min_{1\le j\le m_i} [\mathrm{Re}\, d_j / \delta_j] > 1 \quad (i = 1, 2, \cdots, r)$$

and the conditions given by [8, pp.252-253, (C.4), (C.5) and (C.6)] are also satisfied.

Proof : To establish (2.1), express the H-function in the integrand as [8, pp.251, (C.1)], change the orders of x-integrals and ϕ-integrals, evaluate inner-integrals with the help of [5, p.340], viz.

$$\int_{-\pi/2}^{\pi/2} (\cos x)^{\mu} e^{\nu} dx = \frac{\pi \Gamma(w+v-1)}{2^{w+v-2}\Gamma(w)\Gamma(v)}, \quad \mathrm{Re}(w+v) > 1$$

and use [8, p.251, (C.1)].

Note 1 : The integral (2.1) may be viewed as the several variables analogue of the integral [1, p.88, (2.1)].

3. The Fourier exponential series.

The Fourier exponential series to be established is

$$(3.1)\quad (\cos x_1)^{2[\mu(1)+1]} \cdots (\cos x_r)^{2[\mu(r)+1]} H\begin{bmatrix} z_1(e^{ix_1}\cos x_1)^{t_1} \\ \cdots \\ z_r(e^{ix_r}\cos x_r)^{t_r} \end{bmatrix} =$$

$$= \frac{2^{2r-2(v_1+\cdots+v_r)}}{\Gamma(2v_1)\cdots\Gamma(2v_r)} \sum_{u=-\infty}^{\infty} \cdots \sum_{u=-\infty}^{\infty} 2^{-2i(u_1+\cdots+u_r)}$$

$$e^{-2i\{(u_1x_1+\cdots+u_rx_r)-(v_1x_1+\cdots+v_rx_r)\}}$$

$$H\begin{matrix} 0,n;m_1,n_1+1;\cdots;m_r,n_r+1 \\ p,q;p_1+1,q_1+1;\cdots;p_r+1,q_r+1 \end{matrix}\left[\begin{matrix} z_1 \\ z_r \end{matrix}\middle|\begin{matrix} A;(2-2u_1-2v_1,t_1),C_p;\cdots;(2-2u_r-2v_r,t_r),C_{p_r} \\ B;D_{q_1},(1-2u_1,t_1);\cdots;D_{q_r},(1-2u_r,t_r) \end{matrix}\right]$$

valid under the conditions of (2.1).

Proof : Let

$$(3.2)\quad (\cos x_1)^{2[\mu(1)+1]} \cdots (\cos x_r)^{2[\mu(r)+1]} H\begin{bmatrix} z_1(e^{ix_1}\cos x_1)^{t_1} \\ \\ z_r(e^{ix_r}\cos x_r)^{t_r} \end{bmatrix} =$$

$$= \sum_{u=-\infty}^{\infty} \cdots \sum_{u=-\infty}^{\infty} C_{u_1, \cdots, u_r} \exp[-2i\{(u_1x_1 + \cdots + u_rx_r) - (v_1x_1 + \cdots + v_rx_r)\}]$$

Multiplying both sides of (3.2) by

$$\exp[2i\{(w_1x_1 + \cdots + w_rx_r) - (v_1x_1 + \cdots + v_rx_r)\}]$$

and integrating with respect to $x_1, x_2, \ldots, x_r$, from $-\pi/2$ to $\pi/2$, and using (2.1) and the orthogonality property of the exponential functions [3, p.62]:

$$\int_c^{\infty} \exp\left(\frac{2m\pi ix}{a-b}\right) \exp\left(\frac{2n\pi ix}{a-b}\right) dx = \begin{cases} 0, & m \neq n \\ b-a, & m = n \end{cases}$$

we obtain the value of $C_{u_1, \ldots u_r}$. Sustituting the valued of $C_{u_1, \ldots u_r}$, in (3.2), the Fourier exponential series (3.1) is obtained.

Note 2 : The Fourier exponential series (3.1) may be viewed as the several variables analogue of [1, p.89, (3.1)].

Since on specializing the parameters Fox's H-function of several variables yields almost all special functions appearing in applied mathematics and physical sciences. Therefore, the results presented in this paper are of a general character and hence may encompass several cases of interest.

References

[1]. Bajpai, S.D.: "Some expansion formulae for Fox's H-function involving exponential functions". Proc. Camb. Phil. Soc., 67(1970), 87-92.

[2]. Carslaw, H.S. and Jaeger, J.C.: "Conduction of heat in solids". Clarendon Press. Oxford, 1986.

[3]. Churchill, R. V. : "Fourier series and boundary value problems". McGraw-Hill, New York, 1941.

[4]. Fox, C.: "The G and H-functions as symmetrical Fourier kernels". Trans. Amer. Math. Soc. 98 (1961), 395-429.

[5]. MacRobert, T. M. : "Functions of a complex variable". Macmillan, London, 1962.

[6]. Mathai, A.M. and Saxena, R.K.: "Generalized hypergeometric functions with applications in statistics and physical sciences". Lecture Notes Series No. 348. Springer-Verlag, Berlin, Heidelberg and New York, 1979.

[7]. Mathai, A.M. and Saxena, R.K. : "The H-function with applications in statistics and other disciplines". Wiley Eastern Ltd., New Delhi, 1978.

[8]. Srivastava, H.M., Gupta, K.C. and Goyal, S.P.: "The H-functions of one and two variables with applications". South Asian Publishers, New Delhi, 1982.

Rev. Academia de Ciencias. Zaragoza. 49 (1994)

A CANONICAL REDUCTION OF A CLASS OF PERTURBED TWO-BODY PROBLEMS

L. Floría

Grupo de Mecánica Celeste.
Departamento de Matemática Aplicada a la Ingeniería.
E. T. S. de Ingenieros Industriales. Universidad de Valladolid.
E – 47 011 Valladolid. Spain.

Abstract

In the 8–dimensional, extended phase space of the polar nodal variables we propose a canonical reduction of a Hamiltonian $\mathcal{H}$ that formalizes a wide variety of cases of perturbed two–body motion. Our procedure allows us to formally contract the generic Hamiltonian system under consideration onto a Keplerian one, and develop a simple analytical solution to $\mathcal{H}$ in closed form. In so doing, we solve $\mathcal{H}$ by performing the transition to a set of canonical elements for the dynamical problem linked to $\mathcal{H}$, the independent variable being then proportional to a true–like anomaly. These elements generalize the classical Delaunay–Similar (DS) ones employed by Scheifele and Graf within the framework of the analytical satellite theories in the extended phase space, and contain the perturbation originally present in the Hamiltonian $\mathcal{H}$. As a mere illustration of this approach, some special cases of perturbed two–body motion (formulated by radial intermediaries), borrowed from the Theory of Artificial Earth Satellites, are adduced.

Key words: perturbed two–body orbital motion, polar nodal variables, reducing canonical transformations, generalized Delaunay–Similar (GDS) elements, artificial satellite, radial intermediaries.

AMS (MOS) Subject Classification: 70 F 05, 70 F 15, 70 H 15, 70 M 20, 58 F 05.

PACS Numbers: 95.10.Ce, 95.40.+s, 03.20.+i, 46.10.+z.

1. Introduction

The present paper tackles an analytical investigation concerning a DS–type approach to the canonical reduction of a *general*, homogeneous Hamiltonian $\mathcal{H}$ that formalizes a case of two–body motion in which generic perturbation effects due to certain types of disturbing potentials are taken into account. As special instances after particular choices of the potential, by extending the results obtained in Floría (1991) and paralleling the developments presented in Floría (1993), some radial intermediaries in polar nodal variables (Ferrándiz and Floría, 1991 and 1993) for the so–called *Main Problem* in Artificial Satellite Theory are easily recovered and made to fit into this pattern. To be precise, the general perturbing potential allowed for in $\mathcal{H}$ will contain terms proportional to some negative powers of the radial distance, namely r^{-j} with $j = 0, 1, 2$.

Our procedure allows us to formally contract the generic Hamiltonian system governed by $\mathcal{H}$ onto a conventional Keplerian one. In so doing, we construct a new set of canonical elements for the dynamical problem attached to $\mathcal{H}$. This set constitutes a generalization of the Delaunay–Similar (DS) one (with the true anomaly as the independent variable) applied by Scheifele and Graf (1974) to reduce the aforesaid Main Problem. Other authors (see, e. g., Bond and Broucke 1980, Bond and Janin 1981) also considered this set or some variants of it.

Our elements, obtained without having to seek a complete solution to the Hamilton–Jacobi equation linked to $\mathcal{H}$, are derived by appropriately modifying the technique devised by Deprit (1981a) to perform the transition from polar nodal variables to a set of Scheifele elements, and enjoy the property that they contain the perturbation characterized by the disturbing potentials as an effect incorporated into the generating relations by means of which the new variables are defined.

As a consequence of our approach and the subsequent reduction, a *simple analytical solution* for the general dynamical problem governed by $\mathcal{H}$ is derived in the extended, 8–dimensional phase space. We emphasize that this is achieved by means of a canonical transformation operating on the polar nodal variables, which produces a set of canonical elements of a Delaunay–Similar (DS) type. The transformation is defined from a suitable generating function S of the second type (as dubbed by Goldstein, 1980, Chapter 9) whose functional form is, in principle, inspired by the Hamiltonian $\mathcal{H}$. The development of this transformation requires the evaluation of certain quadratures over the radial variable r, which is performed by a standard procedure with the help of appropriately introduced integration variables of a true–like and eccentric–like anomaly nature, formally interpreted as paralleling the customary picture of a fictitious Keplerian motion characterized by suitably amended elliptic elements.

As the second step in our approach, a proper change of the time parameter in the form of a differential relation connecting the old and the new time variable is considered. The new time variable turns out to be proportional to the perturbed true anomaly pertaining to the aforesaid hypothetic Keplerian motion.

The application of the new canonical variables to $\mathcal{H}$ converts its functional form into the one taken on by the standard Keplerian Hamiltonian when formulated in terms of the aforementioned set of classical DS variables used by Scheifele and Graf (1974). Consequently, the simplified expression of the transformed Hamiltonian also reveals the status of *canonical elements* (in the sense, e. g., of Stiefel and Scheifele, 1971, §18), for the problem originally posed by $\mathcal{H}$, attained by the new variables now introduced.

In particular, a simple analytical solution in our generalized DS (GDS) variables can be easily derived for $\mathcal{H}$. As a result of this construction, we conclude that the solution obtained after this process absorbs the perturbation arising from the potential present in $\mathcal{H}$.

Finally, the Theory of Artificial Satellites supplies us with some examples to which we can apply our general developments. We integrate some radial intermediaries that make up integrable approximations to the J_2 problem of that theory, and find the corresponding analytical solutions in our GDS–approach.

2. Formulation of the Basic Hamiltonian

The extended canonical set of the *Hill–Whittaker polar nodal variables* (Deprit 1981b, §2, pp.113–114), namely $(r\,,\,\theta\,,\,\nu\,,\,t\,;\,p_r\,,\,p_\theta\,,\,p_\nu\,,\,p_0)$, will be considered, where p_0 (the canonical momentum conjugate to the physical time t) is the negative of the total energy in the problem to which the variables are applied (see, e. g. Poincaré 1905, vol. I, Chapter 1, §12, or Stiefel and Scheifele 1971, §30).

In this set of variables the Hamiltonian $\mathcal{H}$ will be formulated as the function

$$\mathcal{H} \;\equiv\; \mathcal{H}_0\,(r\,;\,p_r\,,\,p_\theta\,) \,+\, \sum_{j\,=\,0}^{2} \frac{1}{r^{\,j}}\, V_{j\,,(n_j)}\,(p_\theta\,,\,p_\nu\,,\,p_0\,;\,\varepsilon\,) \,+\, p_0\,,$$

where

$$\mathcal{H}_0\,(r\,;\,p_r\,,\,p_\theta\,) \;=\; \frac{1}{2}\left[\,{p_r}^{\,2} \,+\, \frac{{p_\theta}^{\,2}}{r^{\,2}}\,\right] \,-\, \frac{\mu}{r}$$

stands for the conventional Keplerian Hamiltonian, and

$$V_{j\,,(n_j)} \,\equiv\, V_{j\,,(n_j)}\,(p_\theta\,,\,p_\nu\,,\,p_0\,;\,\varepsilon) \;=\; \sum_{l\,=\,1}^{n_j} \varepsilon^{\,l}\, \mathcal{V}_{j\,,l}\,(p_\theta\,,\,p_\nu\,,\,p_0\,)$$

formalizes a perturbing potential acting on $\mathcal{H}_0$, and will be an expansion (truncated at the higher order n_j) in ascending powers of a small adimensional parameter ε, the coefficients

$\mathcal{V}_{j,l}(p_\theta\,,\,p_\nu\,,\,p_0)$ of such an expansion being functions of the canonical momenta p_θ, p_ν and p_0. In what follows, the specific dependence of the $V_{j,(n_j)}$ on these momenta and on ε will not be significant.

For the sake of simplicity in the notations, from now on the subscripts (n_j) will be left out, unless such omissions and the subsequent simplification of the notations could cause misunderstanding.

The solution to the problem given by $\mathcal{H}$ will be approached by means of the construction of a canonical transformation, and the resulting Hamiltonian will be integrated in the new variables.

3. Development of the Transformation

With the aim of performing a canonical reduction of this Hamiltonian and deriving a solution to it, we propose the change of phase variables to a new set of generalized DS (GDS, for short) variables,

$$(r\,,\,\theta\,,\,\nu\,,\,t\,;\,p_r\,,\,p_\theta\,,\,p_\nu\,,\,p_0)\ \xrightarrow{\ S\ }\ (q_\Phi\,,\,q_L\,,\,q_G\,,\,q_N\,;\,\Phi\,,\,L\,,\,G\,,\,N)\,,$$

implicitly defined via a generating function depending on the old coordinates and the new momenta

$$S \equiv S(r\,,\theta\,,\nu\,,t\,;\Phi\,,\,L\,,\,G\,,\,N) = \theta\,G + \nu\,N + t\,L + \int_{r_0}^{r} \sqrt{Q}\ dr\,,$$

the function under the radical sign being

$$Q \equiv Q(r\,;\Phi\,,\,L\,,\,G\,,\,N\,;\varepsilon) = a_0 + \frac{a_1}{r} + \frac{a_2}{r^2} = \frac{1}{r^2}\left[a_0 r^2 + a_1 r + a_2\right],$$

having introduced the abbreviations

$$a_0 = -\,2L - 2V_0(G\,,\,N\,,\,L\,;\,\varepsilon)\,,$$

$$a_1 = 2\mu - 2V_1(G\,,\,N\,,\,L\,;\,\varepsilon)\,,$$

$$a_2 = -\,\gamma^2 - 2V_2(G\,,\,N\,,\,L\,;\,\varepsilon)\,,$$

and γ designates the function of the new canonical momenta given by

$$\gamma \equiv \gamma(\Phi\,,\,L\,,\,G) = G - \Phi + \frac{\mu}{\sqrt{2L}}\,.$$

Consequently

$$\sqrt{Q} = \sqrt{-2(L + V_0) + \frac{2(\mu - V_1)}{r} - \frac{(\gamma^2 + 2V_2)}{r^2}}\ ;$$

for the sake of conciseness, the functional dependence of the V_j on the new momenta has been omitted. Understand also that the lower limit r_0 in the integral is any simple zero of the function Q, that, in particular, can be chosen as the lowest positive root of the r–equation $Q(r\,;\,\Phi\,,\,G\,,\,N\,,\,L\,;\,\varepsilon) \;=\; 0\,.$

The transition to the new variables is performed through the generating relations derived from S, which yields the set of implicit transformation formulae

$$
\begin{aligned}
p_r \;&\equiv\; \frac{dr}{dt} \;=\; \frac{\partial S}{\partial r} \;=\; \sqrt{Q}\,,\\
p_\theta \;&=\; \frac{\partial S}{\partial \theta} \;=\; G\,,\\
p_\nu \;&=\; \frac{\partial S}{\partial \nu} \;=\; N\,,\\
p_0 \;&=\; \frac{\partial S}{\partial t} \;=\; L\,,\\
q_\Phi \;&=\; \frac{\partial S}{\partial \Phi} \;=\; \gamma\, I_2\,,\\
q_G \;&=\; \frac{\partial S}{\partial G} \;=\; \theta \;-\; \frac{\partial V_0}{\partial G}\, I_0 \;-\; \frac{\partial V_1}{\partial G}\, I_1 \;-\; \left[\,\gamma \;+\; \frac{\partial V_2}{\partial G}\,\right] I_2\,,\\
q_N \;&=\; \frac{\partial S}{\partial N} \;=\; \nu \;-\; \frac{\partial V_0}{\partial N}\, I_0 \;-\; \frac{\partial V_1}{\partial N}\, I_1 \;-\; \frac{\partial V_2}{\partial N}\, I_2\,,\\
q_L \;&=\; \frac{\partial S}{\partial L} \;=\; t \;-\; \left[\,1 \;+\; \frac{\partial V_0}{\partial L}\,\right] I_0 \;-\; \frac{\partial V_1}{\partial L}\, I_1 \;-\; \left[\,-\,\gamma\,\frac{\mu}{(2\,L)^{3/2}} \;+\; \frac{\partial V_2}{\partial L}\,\right] I_2\,,
\end{aligned}
$$

where

$$
I_m \;=\; \int_{r_0}^{r} \frac{dr}{r^m\,\sqrt{Q}}\,, \qquad m \;=\; 0\,,\,1\,,\,2\,.
$$

The equation for q_L can be looked on as a general expression for *Kepler's equation.*

The way of proceeding to evaluate these quadratures is based on the idea of adapting a technique (classically applied to a pure Kepler problem to derive Delaunay variables as done, e.g., in Deprit 1981b, pp. 115–118, and for quasi–Keplerian systems in the same article Deprit 1981b, pp. 124–126) to the considered, extended Hamiltonian $\mathcal{H}$ and taking advantage of the homogeneous canonical formalism. In this respect we also refer to Deprit (1981a), whose procedure is modified so as to take into account the non–Keplerian terms present in the potential of $\mathcal{H}$.

The introduction of a set of appropriate subsidiary quantities a, e, p, n, Γ, μ^*,

depending on the new canonical momenta, by means of the formulae

$$a = \frac{\mu^*}{2(L + V_0)}, \qquad \mu^* = \mu - V_1, \qquad \mu^* = n^2 a^3,$$

$$\gamma^2 + 2V_2 = \mu^* a(1 - e^2) \equiv \Gamma^2, \qquad p = a(1 - e^2) = \frac{\Gamma^2}{\mu^*},$$

$$e^2 = 1 - \frac{\Gamma^2}{\mu^* a} = 1 - \frac{2(L + V_0)\Gamma^2}{(\mu^*)^2},$$

allows one to factorize Q as the product

$$Q = \frac{2\mu^*}{r} - \frac{\mu^*}{a} - \frac{\mu^* a(1 - e^2)}{r^2} = \frac{\mu^*}{a}\left[\frac{a(1 + e)}{r} - 1\right]\left[1 - \frac{a(1 - e)}{r}\right].$$

These formulae resemble those formally holding for a hypothetic, Keplerian motion characterized by the above *elliptic elements* (a, e, p) with Γ as the *modified angular momentum magnitude*, provided that μ^* is taken as the gravitational parameter. Correspondingly, bearing in mind that p_θ, p_ν and p_0 are not changed by the transformation, and translating the above expressions into the respective ones in terms of the original polar nodal variables, the moving mass can be regarded as *simulating a Kepler motion controlled by the fictitious Hamiltonian*

$$\mathcal{H}^* = \frac{1}{2}\left[p_r{}^2 + \frac{\Gamma^2}{r^2}\right] - \frac{\mu^*}{r}.$$

Our development assumes that the quantity e is such that $0 \leq e < 1$, in which case the roots of the equation $Q(r; \Phi, L, G, N; \varepsilon) = 0$ are

$$0 < r_0 \equiv r_p = a(1 - e) \leq r_1 \equiv r_a = a(1 + e),$$

interpreted as the perturbed pericentre and apocentre radial distances in terms of the Keplerian–like quantities a and e. Consequently

$$Q = \frac{\mu^*}{ar^2}\{a(1 + e) - r\}\{r - a(1 - e)\} = \frac{\mu^* a e^2}{r^2}\left[1 - \left(\frac{a - r}{ae}\right)^2\right].$$

The *auxiliary integration variables* E, of the eccentric–anomaly–type, and f, of the true–anomaly–type, defined by

$$r = a(1 - e\cos E), \qquad \frac{1}{r} = \frac{1 + e\cos f}{p},$$

yield the following expressions for Q:

$$Q = \frac{\mu^* e^2}{a}\frac{\sin^2 E}{(1 - e\cos E)^2}, \qquad Q = \frac{\mu^*}{p} e^2 \sin^2 f,$$

whence the quadratures I_0 and I_2 are found to result in

$$I_0 = \sqrt{\frac{a^3}{\mu^*}}\ (E - e\sin E)\,, \qquad I_2 = \frac{f}{\sqrt{\mu^* p}} = \frac{f}{\Gamma}\,.$$

In this way, these auxiliary variables can respectively be contemplated as the *eccentric and true anomaly along the aforesaid fictitious Keplerian motion.*

The quadrature I_1 is performed with the help of the variables E and f. Putting

$$I_1 = \int_{r_0}^{r} \frac{r\,dr}{r^2\sqrt{Q}} = -\int_{r_0}^{r} \frac{r\,d(1/r)}{\sqrt{Q}} = \sqrt{\frac{p}{\mu^*}} \int_0^f \frac{df}{1 + e\cos f}\,,$$

and, after introducing the quantity $\eta = \sqrt{1 - e^2}$, the last integral is calculated by means of the Keplerian–like relations

$$r\cos f = a(\cos E - e)\,, \qquad r\sin f = a\eta\sin E\,, \qquad r = a(1 - e\cos E)\,,$$

which allows one to express $\cos f$ and df in terms of the variable E in such a way that

$$\int_0^f \frac{df}{1 + e\cos f} = \int_0^E \frac{dE}{\eta} = \frac{E}{\eta} \implies I_1 = \sqrt{\frac{p}{\mu^*}}\,\frac{E}{\eta} = \frac{\Gamma}{\mu^*\eta}\,E\,.$$

The preceding preliminary relations allow one to complete the set of transformation formulae, namely:

$$q_\Phi = \frac{\gamma}{\Gamma}\,f\,,$$

$$q_G = \theta - \frac{\partial V_0}{\partial G}\,\frac{1}{n}\,(E - e\sin E) - \frac{\partial V_1}{\partial G}\,\frac{\Gamma E}{\mu^*\eta} - \left[\gamma + \frac{\partial V_2}{\partial G}\right]\frac{f}{\Gamma}\,,$$

$$q_N = \nu - \frac{\partial V_0}{\partial N}\,\frac{1}{n}\,(E - e\sin E) - \frac{\partial V_1}{\partial N}\,\frac{\Gamma E}{\mu^*\eta} - \frac{\partial V_2}{\partial N}\,\frac{f}{\Gamma}\,,$$

$$q_L = t - \left[1 + \frac{\partial V_0}{\partial L}\right]\frac{1}{n}\,(E - e\sin E) - \frac{\partial V_1}{\partial L}\,\frac{\Gamma E}{\mu^*\eta} - \left[\frac{\partial V_2}{\partial L} - \frac{\mu\gamma}{(2L)^{3/2}}\right]\frac{f}{\Gamma}\,.$$

In the next section the canonical transformation here obtained will be applied to the Hamiltonian $\mathcal{H}$, after which it will be seen that the functional form of the new Hamiltonian is substantially simplified if an adequate change of the independent variable is performed.

4. Transformed Hamiltonian. New Time Parameter

The Hamiltonian $\mathcal{H}$ will now be reduced by the effect of the preceding transformation, which maps it onto a function $\widetilde{\mathcal{H}}$ admitting a simple factorized form:

$$\mathcal{H}(r, -, -, -\,;\, p_r, p_\theta, p_\nu, p_0) \longrightarrow \widetilde{\mathcal{H}} = \frac{1}{2r^2}\left(G^2 - \gamma^2\right),$$

where r must be regarded as a function of the GDS canonical variables through the Keplerian–like relations defining E and f.

The homogeneous formalism in the extended phase space makes possible the introduction of *new independent variables* other than the physical time t in a rather simple way. A very common device for doing this consists of considering suitable time transformations $t \longrightarrow \tau$ of the *Sundman type*, the new fictitious time τ being defined by a differential relation of the form $dt = \tilde{f}\, d\tau$, where the function $\tilde{f}$ is taken to be proportional to a power of the radial distance r through a coefficient which is a constant or a function of the new canonical variables, say $\tilde{f} = k\, r^{\alpha}$.

In the present case, the choice of $\tilde{f}$ as the function

$$\tilde{f} \;=\; 2\, r^2/(G \,+\, \gamma)\,,$$

proportional to r^2, changes the independent variable from physical time t to a fictitious time τ that can be considered as proportional to a "*generalized true anomaly*" along the fictitious Keplerian motion previously mentioned.

This reparametrizing transformation is reflected in a significant change of expression in the Hamiltonian: the new Hamiltonian corresponding to τ as the independent variable is readily found to be

$$\mathcal{K} \;=\; \widetilde{\mathcal{H}}\,\tilde{f} \;=\; G \,-\, \gamma \;\Longrightarrow\; \mathcal{K} \;=\; \Phi \,-\, \frac{\mu}{\sqrt{2L}}\;,$$

whose structure is easily recognizable as formulating a pure Kepler motion in terms of the DS canonical variables presented by Scheifele and Graf (1974), page 3, Remark(b), Bond and Broucke (1980), page 359, or Bond and Janin (1981), page 161. The canonical equations of motion derived from the reduced Hamiltonian $\mathcal{K}$ are immediately integrated, yielding *a simple parametrical solution* with τ as the independent variable:

$$q_{\Phi} \;=\; \tau \,+\, \text{const.} \quad \text{and} \quad q_L \;=\; \frac{\mu}{(2L)^{3/2}}\,\tau \,+\, \text{const.}$$

are linear functions of the fictitious time τ, and the remaining GDS variables behave like constants of the motion. It is seen that the canonical coordinate q_{Φ} is the pseudo–time τ up to a constant.

The nature of the preceding canonical solution shows that these GDS variables can be regarded as making up *a set of canonical elements* (in the sense, e. g., of Stiefel and Scheifele 1971, Section 18) when applied to the Hamiltonian $\mathcal{H}$, provided that the pseudo–time τ is used as the independent variable.

Since $\mathcal{K}$ does not depend explicitly on the perturbation parameter ε, the proposed new set of variables *does* intrinsically contain all variations due to the potential associate to $\mathcal{H}$.

To sum up: a reduction of $\mathcal{H}$ to the Hamiltonian $\mathcal{K}$ corresponding to a hypothetic conventional Keplerian one has been performed, and the proposed variables absorb the perturbing influence originally included in the potential of $\mathcal{H}$.

As for the way of obtaining the physical time t, observe that its determination does not require the integration of the differential relation $dt = \tilde{f}\, d\tau$. As in the case of the classical DS elements (Scheifele and Graf 1974, page 3, Remark(a); Bond and Janin 1981, page 159), an analogous remark is now pertinent: t is obtained from τ via the time element q_L by means of the *generalized Kepler equation*

$$t = q_L + \left[1 + \frac{\partial V_0}{\partial L}\right]\frac{1}{n}\,(E - e\sin E) + \frac{\partial V_1}{\partial L}\,\frac{\Gamma E}{\mu^* \eta} + \left[\frac{\partial V_2}{\partial L} - \frac{\mu\gamma}{(2L)^{3/2}}\right]\frac{f}{\Gamma},$$

taking into account the expressions for the remaining elements as obtained from the canonical solution to the reduced Hamiltonian $\mathcal{K}$.

By making $\varepsilon = 0$ in the generating function S, the resulting canonical transformation performs the transition from Hill–Whittaker variables to the set of canonical Delaunay–Similar variables used by Scheifele and Graf (1974).

Remember that the use of sets of polar–like variables and other ones derived from them introduces *virtual singularities* due to small values of eccentricity and/or values of the inclination close to 0° or 180°. As done in Floría and Ferrándiz (1991), associate canonical sets of *generalized Poincaré–Similar (GPS) variables* corresponding to the preceding GDS ones can be defined. Such GPS variables also incorporate the contribution of the perturbing potential into their definition, and provide a way of avoiding the appearance of the above–mentioned virtual singularities when studies of perturbations are carried out.

5. Form of the Solution

In the light of the preceding considerations, from the transformation equations and after solving for the original polar nodal variables in terms of the auxiliary integration variables E and f and taking advantage of the subsidiary Keplerian–kind quantities previously introduced, a Keplerian–like solution to $\mathcal{H}$ can be set up by means of a parametric

representation schematized in the form

$$r = a(1 - e\cos E) = \frac{p}{1 + e\cos f},$$

$$p_r = \sqrt{\frac{\mu^*}{a}}\,\frac{e\sin E}{(1 - e\cos E)} = \sqrt{\frac{\mu^*}{p}}\; e\sin f,$$

$$\theta = q_G + \frac{\partial V_0}{\partial G}\frac{1}{n}(E - e\sin E) + \frac{\partial V_1}{\partial G}\frac{\Gamma E}{\mu^*\eta} + \left[\gamma + \frac{\partial V_2}{\partial G}\right]\frac{f}{\Gamma},$$

$$\nu = q_N + \frac{\partial V_0}{\partial N}\frac{1}{n}(E - e\sin E) + \frac{\partial V_1}{\partial N}\frac{\Gamma E}{\mu^*\eta} + \frac{\partial V_2}{\partial N}\frac{f}{\Gamma},$$

$$t = q_L + \left[1 + \frac{\partial V_0}{\partial L}\right]\frac{1}{n}(E - e\sin E) + \frac{\partial V_1}{\partial L}\frac{\Gamma E}{\mu^*\eta} + \left[\frac{\partial V_2}{\partial L} - \frac{\mu\gamma}{(2L)^{3/2}}\right]\frac{f}{\Gamma},$$

$$p_\theta = G = \text{const.}, \qquad p_\nu = N = \text{const.}, \qquad p_0 = L = \text{const.},$$

together with

$$q_\Phi = \frac{\gamma}{\Gamma}f = \tau + \text{const.}, \qquad q_G = \text{const.},$$

$$q_L = \frac{\mu}{(2L)^{3/2}}\tau + \text{const.}, \qquad q_N = \text{const.}$$

The generality of this pattern facilitates a compact and unified treatment of a wide class of perturbed two–body problems in GDS variables. To illustrate this approach, the next Section contains, as special cases under specific choices of the perturbing potential that distorts the purely Keplerian orbit, the corresponding analytical solutions to some radial intermediaries for the J_2 problem of the theory of artificial Earth satellites.

6. Solution to Some Radial Intermediaries

The potentials considered in this section exemplify some remarkable particular perturbations that only contain a power of r with the exponent -2 (Deprit's 1981b intermediary), -1 or 0 (Ferrándiz and Floría 1993). As stated in Ferrándiz and Floría (1993), the new intermediaries share, with that introduced by Deprit, a great formal proximity to the Keplerian picture of motion. This statement is, once again, confirmed here by the present analysis, in view of the simple solution in closed form (see below) that they admit.

The potentials will first be expressed in extended Hill–Whittaker variables, and the following remarks concerning the notations are in order: the symbol $\varepsilon = -J_2$, which denotes the dimensionless oblateness parameter of the primary, will act as the (small) perturbation parameter; μ stands for the gravitational constant of the central body, and R_e represents its mean equatorial radius. Remember also that the functions of the inclination

c and s are given by

$$c \equiv \cos I = \frac{p_\nu}{p_\theta}, \qquad s \equiv \sin I, \qquad s^2 = 1 - c^2 .$$

For convenience, the expression for the potential will be given in terms of powers of c^2, instead of doing it in terms of powers of s^2.

Since the canonical momenta p_θ, p_ν, and p_0 remain unchanged under the considered transformations, there is no difficulty in translating partial derivatives with respect to them into partial derivatives with respect to the corresponding new momenta, a mere change of notation being sufficient.

6.1 Deprit's Intermediary (1981b)

The potential yielding this intermediary is obtained for

$$V_0(p_\theta, p_\nu, p_0; \varepsilon) \equiv 0, \qquad V_1(p_\theta, p_\nu, p_0; \varepsilon) \equiv 0,$$
$$V_2 \equiv V_{2,(1)}^{(D)}(p_\theta, p_\nu, -; \varepsilon) = \varepsilon \frac{\mu^2 {R_e}^2}{4 {p_\theta}^2} (3c^2 - 1),$$

and therefore

$$\frac{\partial V_2}{\partial G} = \varepsilon \frac{\mu^2 {R_e}^2}{2 G^3} \left[1 - \frac{6 N^2}{G^2} \right], \qquad \frac{\partial V_2}{\partial N} = \varepsilon \frac{3 \mu^2 {R_e}^2 N}{2 G^4},$$

while the remaining partial derivatives involved in the transformation vanish identically.

Auxiliary quantities:

$$\mu^* \equiv \mu, \qquad a = \frac{\mu}{2L}, \qquad n = \frac{(2L)^{3/2}}{\mu},$$
$$\Gamma^2 \equiv \gamma^2 + 2V_2 = \mu a (1 - e^2), \qquad p = a(1 - e^2) = \frac{\Gamma^2}{\mu},$$
$$e^2 = 1 - \frac{\Gamma^2}{\mu a} = 1 - \frac{2L\Gamma^2}{\mu^2}, \qquad q_\Phi = \frac{\gamma}{\Gamma} f .$$

Fictitious Keplerian–like Hamiltonian:

$$\mathcal{H}^* = \frac{1}{2} \left[{p_r}^2 + \frac{\Gamma^2}{r^2} \right] - \frac{\mu}{r} .$$

Solution to the intermediary:

$$r = a(1 - e\cos E) = \frac{p}{1 + e\cos f},$$

$$p_r = \sqrt{\frac{\mu}{a}}\,\frac{e\,\sin E}{(1 - e\cos E)} = \sqrt{\frac{\mu}{p}}\; e\,\sin f,$$

$$\theta = q_G + \left[\gamma + \frac{\partial V_2}{\partial G}\right]\frac{f}{\Gamma} = q_G + \frac{\partial \Gamma}{\partial G}\,f,$$

$$\nu = q_N + \frac{\partial V_2}{\partial N}\,\frac{f}{\Gamma} = q_N + \frac{\partial \Gamma}{\partial N}\,f,$$

$$t = q_L + \frac{1}{n}(E - e\,\sin E) - \frac{\mu\gamma}{(2L)^{3/2}}\,\frac{f}{\Gamma},$$

$$p_\theta = G = \text{const.}, \qquad p_\nu = N = \text{const.}, \qquad p_0 = L = \text{const.}$$

Notice that, in the present case, the *generalized Kepler equation* can also be written in the more familiar form

$$t = q_L + \frac{\mu}{(2L)^{3/2}}\,(E - e\,\sin E - q_\Phi).$$

6.2 A Brouwer-like Intermediary (Ferrándiz, 1990)

This intermediary, directly developed by Ferrándiz in the extended phase space and presented in Ferrándiz and Floría (1993), contains a first order contribution emanating from a potential V_0 of the form

$$V_{0,(1)} \equiv V_{0,(1)}(p_\theta, p_\nu, p_0\,;\,\varepsilon) = \varepsilon\,\mathcal{V}_{0,1}$$

$$= \varepsilon\,\frac{\mu\,{R_e}^2}{4\,{p_\theta}^3}\,(3c^2 - 1)\,(2p_0)^{3/2}.$$

The corresponding Hamiltonian is close to the secular Hamiltonian of Brouwer's solution (1959), although not exactly the same, due to a different treatment of some quantities involved in the derivation. Further details can be found in the aforementioned paper by Ferrándiz and Floría (1993).

Because of a different choice of the perturbation parameter, this expression is slightly different from that given in the said paper: on that occasion, the choice was $\varepsilon = -\mu\,{R_e}^2\,J_2$, while now $\varepsilon = -J_2$.

The list of partial derivatives involved in the transformation will now be

$$\frac{\partial V_0}{\partial G} = \varepsilon \frac{3\mu {R_e}^2}{4G^4}\left[1 - \frac{5N^2}{G^2}\right](2L)^{3/2},$$

$$\frac{\partial V_0}{\partial N} = \varepsilon \frac{3\mu {R_e}^2 N}{2G^5}(2L)^{3/2},$$

$$\frac{\partial V_0}{\partial L} = \varepsilon \frac{3\mu {R_e}^2}{4G^3}\left[\frac{3N^2}{G^2} - 1\right](2L)^{1/2}.$$

Auxiliary quantities:

$$\mu^* \equiv \mu, \qquad a = \frac{\mu}{2(L + V_0)}, \qquad n = \frac{(2L + V_0)^{3/2}}{\mu},$$

$$\Gamma^2 \equiv \gamma^2 = \mu a(1 - e^2), \qquad p = a(1 - e^2) = \frac{\gamma^2}{\mu},$$

$$e^2 = 1 - \frac{\gamma^2}{\mu a} = 1 - \frac{2(L + V_0)\gamma^2}{\mu^2}, \qquad q_\Phi = f.$$

Fictitious Keplerian–like Hamiltonian:

$$\mathcal{H}^* = \frac{1}{2}\left[{p_r}^2 + \frac{{p_\theta}^2}{r^2}\right] - \frac{\mu}{r}.$$

Solution to the intermediary:

$$r = a(1 - e\cos E) = \frac{p}{1 + e\cos f},$$

$$p_r = \sqrt{\frac{\mu}{a}}\,\frac{e\sin E}{(1 - e\cos E)} = \sqrt{\frac{\mu}{p}}\; e\sin f,$$

$$\theta = q_G + \frac{\partial V_0}{\partial G}\frac{1}{n}(E - e\sin E) + f,$$

$$\nu = q_N + \frac{\partial V_0}{\partial N}\frac{1}{n}(E - e\sin E),$$

$$t = q_L + \left[1 + \frac{\partial V_0}{\partial L}\right]\frac{1}{n}(E - e\sin E) - \frac{\mu}{(2L)^{3/2}}f,$$

$$p_\theta = G = \text{const.}, \qquad p_\nu = N = \text{const.}, \qquad p_0 = L = \text{const.}$$

Due to the relation $q_\Phi = f$, the true–like anomaly has become a canonical variable in the new set of generalized DS elements.

6.3 Ferrándiz' Intermediary (1990)

This is an r^{-1} – radial intermediary and was also presented in Ferrándiz and Floría (1993); remembering the above remark concerning the perturbation parameter, its functional form is

$$V_{1,(1)} \equiv V_{1,(1)}(p_\theta, p_\nu, p_0; \varepsilon) = \varepsilon \mathcal{V}_{1,1}$$

$$= \varepsilon \frac{\mu^2 {R_e}^2}{4 {p_\theta}^3} (3c^2 - 1) \sqrt{2p_0},$$

and the derivatives

$$\frac{\partial V_1}{\partial G} = \varepsilon \frac{3\mu^2 {R_e}^2}{4G^4} \left[1 - \frac{5N^2}{G^2}\right] \sqrt{2L},$$

$$\frac{\partial V_1}{\partial N} = \varepsilon \frac{3\mu^2 {R_e}^2 N}{2G^5} \sqrt{2L},$$

$$\frac{\partial V_1}{\partial L} = \varepsilon \frac{\mu^2 {R_e}^2}{4G^3} \left[\frac{3N^2}{G^2} - 1\right] (2L)^{-1/2}.$$

Auxiliary quantities:

$$\mu^* = \mu - V_1, \qquad a = \frac{\mu^*}{2L}, \qquad n = \frac{(2L)^{3/2}}{\mu^*},$$

$$\Gamma^2 \equiv \gamma^2 = \mu^* a (1 - e^2), \qquad p = a(1 - e^2) = \frac{\gamma^2}{\mu^*},$$

$$e^2 = 1 - \frac{\gamma^2}{\mu^* a} = 1 - \frac{2L\gamma^2}{(\mu^*)^2}, \qquad q_\Phi = f.$$

Fictitious Keplerian–like Hamiltonian:

$$\mathcal{H}^* = \frac{1}{2}\left[{p_r}^2 + \frac{p_\theta^2}{r^2}\right] - \frac{\mu^*}{r}.$$

Solution to the intermediary:

$$r = a(1 - e\cos E) = \frac{p}{1 + e\cos f},$$

$$p_r = \sqrt{\frac{\mu^*}{a}}\ \frac{e\sin E}{(1 - e\cos E)} = \sqrt{\frac{\mu^*}{p}}\ e\sin f,$$

$$\theta = q_G + \frac{\partial V_1}{\partial G}\ \frac{\gamma E}{\mu^* \eta} + f$$

$$\nu = q_N + \frac{\partial V_1}{\partial N}\ \frac{\gamma E}{\mu^* \eta},$$

$$t = q_L + \frac{1}{n}\,(E - e\sin E) + \frac{\partial V_1}{\partial L}\ \frac{\gamma E}{\mu^* \eta} - \frac{\mu}{(2L)^{3/2}}\, f,$$

$$p_\theta = G = \text{const.}, \qquad p_\nu = N = \text{const.}, \qquad p_0 = L = \text{const.}$$

In this case, the auxiliary variable f coincides with the canonical variable q_Φ.

Acknowledgements

This work has been partially supported by CICYT, Comisión Interministerial de Ciencia y Tecnología of Spain, within the National Programme of Space Research under Project ESP. 93–741.

References

[1] Bond, V. and Broucke, R.: 1980, 'Analytical Satellite Theory in Extended Phase Space', *Celestial Mechanics* **21**, 357–360.

[2] Bond, V. R. and Janin, G.: 1981, 'Canonical Orbital Elements in Terms of an Arbitrary Independent Variable', *Celestial Mechanics* **23**, 159–172.

[3] Brouwer, D.: 1959, 'Solution of the Problem of Artificial Satellite Theory without Drag', *Astronomical Journal* **64**, 378–397.

[4] Deprit, A.: 1981a, 'A Note Concerning the TR–Transformation', *Celestial Mechanics* **23**, 299–305.

[5] Deprit, A.: 1981b, 'The Elimination of the Parallax in Satellite Theory', *Celestial Mechanics* **24**, 111–153.

[6] Ferrándiz, J. M. and Floría, L.: 1991, 'Towards a Systematic Definition of Intermediaries in the Theory of Artificial Satellites', *Bulletin of the Astronomical Institutes of Czechoslovakia* **42**, 401–407.

[7] Ferrándiz, J. M. and Floría, L.: 1993, 'New Intermediaries for the Main Problem in Satellite Theory'. In: K. B. Bhatnagar (Editor), *Instability, Chaos and Predictability in Celestial Mechanics and Stellar Dynamics*, 341–352. Nova Science Publishers, New York.

[8] Floría, L.: 1991, 'Generalized DS Elements of an Arbitrary Order', *Boletín Astronómico del Observatorio de Madrid* **12**, 68–77.

[9] Floría, L.: 1993, 'Canonical Elements and Keplerian–like Solutions for Intermediary Orbits of Satellites of an Oblate Planet', *Celestial Mechanics and Dynamical Astronomy* **57**, 203–223.

[10] Floría, L. and Ferrándiz, J. M.: 1991, 'Poincaré–Similar Variables Including J_2 Secular Effects'. In: A. E. Roy (Editor), *Predictability, Stability and Chaos in N–Body Dynamical Systems*, NATO ASI Series **B**, vol. **272**, 297–303. Plenum Publishing Corporation, New York.

[11] Goldstein, H.: 1980, *Classical Mechanics (Second Edition)*, Addison Wesley.

[12] Poincaré, H.: 1905, *Leçons de Mécanique Céleste (professées à la Sorbonne)*, vol. I. (Théorie générale des perturbations planetaires), Gauthier–Villars, Paris.

[13] Scheifele, G. and Graf, O.: 1974, 'Analytical Satellite Theories Based on a New Set of Canonical Elements', AIAA Mechanics and Control of Flight Conference, AIAA Paper No. 74–838, Anaheim, California.

[14] Stiefel, E. L. and Scheifele, G.: 1971, *Linear and Regular Celestial Mechanics*, Springer.

Rev. Academia de Ciencias. Zaragoza. **49** *(1994)*

ON CERTAIN PARTIAL DERIVATIVES INVOLVED IN A DELAUNAY NORMALIZATION PROCESS

J. Alvarez and L. Floría

Grupo de Mecánica Celeste.
Dept. de Matemática Aplicada a la Ingeniería.
E. T. S. I. I. Universidad de Valladolid.
E – 47 011 Valladolid. Spain.

Abstract

This paper concerns the calculation of certain auxiliary partial derivatives required to reduce perturbed Keplerian systems to Delaunay normal form, at least at the first order. To this end, we elucidate the way in which the functions involved in the intermediate reckoning work depend on the dynamical variables of interest, and the detail of some elusive steps is thoroughly considered. To be precise, we have in mind the case of the hyperbolic–type orbital motion of an artificial satellite of an oblate planet, but applications to other dynamical systems can be found.

Key words: perturbed Keplerian systems, Delaunay normalization, hyperbolic–type orbital motion, artificial satellite, oblateness perturbation.

AMS (MOS) Subject Classification: 70 F 05, 70 F 15, 70 H 15, 70 M 20, 58 F 05.

PACS Numbers: 95.10.Ce, 95.40.+s, 03.20.+i, 46.10.+z.

1. Introduction

The present note was originally motivated by some developments due to Brouwer (1959) and Hori (1961). These authors worked out the analytical treatment of the oblateness perturbation problem in Satellite Theory by using the canonical perturbation method attributed to von Zeipel. As well known, this method mixes the coordinates and the momenta belonging to different sets of canonical variables, since the generating function of the transformation at issue depends on both old and new variables. A summary of the

main contents of Brouwer's article can also be found, e. g., in Brouwer and Clemence (1961), Chapter XVII, §12.

Remember that Brouwer approached the *Main Problem* of Artificial Satellite Theory by formulating it in Delaunay variables and seeking an approximate canonical solution to it. His way of analytically solving the corresponding Hamiltonian resorts to the construction of a near–identity canonical transformation creating new ignorable coordinates of a similar type. The said transformation being defined by a generating function, he obtained the generator by means of the Poincaré–von Zeipel perturbation method. After solving the averaged Hamiltonian, the determination of the solution to the original problem requires the knowledge of the periodic perturbations, which are calculated through the partial derivatives of the aforementioned generating function.

Hori (1961), in his investigation of the *hyperbolic* motion of an artificial satellite acted upon by the potential characterizing the *Main Problem* in the Zonal Satellite Theory, formulated the J_2 perturbation problem in terms of a variant of the classical Delaunay elements that is applicable to the study of hyperbolic orbital motion. He derived this set by adapting the construction that, via the Whittaker method, Brouwer and Clemence (1961), Chapter XI, §4 and §9, had offered in the context of elliptic motion. Next, he also obtained a first–order analytical solution by means of a near–identity canonical transformation whose generating function was determined by devising an appropriate modification of the Poincaré–von Zeipel procedure. He replaced the usual periodicity conditions by the device of imposing conditions at $r = \infty$.

Contemplated within the general framework of the problem of motion of an artificial Earth satellite, the treatment carried out by Hori is intended as a translation (at least up to the first order) of Brouwer's solution (1959) to the case of hyperbolic–like orbits. The analytical tools applied by Hori are essentially of the same nature as those previously employed by Brouwer.

In the present paper, the *Main Problem* of the Artificial Satellite Theory will be considered, also under conditions such that the problem results in the case of orbital motion with a positive value for the total energy. After formulating the Hamiltonian in a set of Delaunay variables applicable to *hyperbolic–like orbits*, we can solve this case of the J_2 problem by means of a variant of the *Lie transform technique* (Hori, 1966; Deprit, 1969), which allows us to perform a near–identity canonical transformation to a new set of variables producing a Delaunay normalization of our perturbed dynamical system. This process is very similar, and parallel, to that carried out by Hori (1961). For this reason we shall not go into details of the required intermediate developments, but we shall report on some elusive calculations whose particulars we have never found in the literature and

could turn out to be somewhat misleading, since there can be some muddle over the explicit and implicit functional dependence of certain expressions involved in the calculation.

2. Review of the Delaunay Variables in Hyperbolic Motion

Following Deprit (1982), p. 9, the *Delaunay normal form* of a perturbed Keplerian system is obtained in two steps: (i) the expansion of the original Hamiltonian in Delaunay variables; and (ii) the elimination of the mean anomaly from the transformed Hamiltonian.

To achieve the first step, use will be made of a canonical set of Delaunay variables which is applicable to hyperbolic motion. The set derived by Hori (1961) from the Keplerian orbital elements via the Whittaker method was obtained by Floría (1990; and 1993, §9.2) in appropriately modifying the *Delaunay map* considered by Deprit (1981), §2, pp.114–118, for the treatment of elliptic motion. This Delaunay mapping operates on the phase space of the Hill–Whittaker polar nodal variables, denoted by the symbols $(\, r\,,\, \theta\,,\, \nu\;;\; p_r\,,\, p_\theta\,,\, p_\nu\,)$, and allows us to construct a set of Delaunay elements $(\, l\,,\, g\,,\, h\;;\; L\,,\, G\,,\, H\,)$.

The interpretation of the polar nodal variables is the following: r stands for the radial distance from the primary's centre of mass to the moving point; θ is the argument of latitude of the orbiter, measured from the ascending node; ν designates the argument of longitude of the ascending node. As for the canonical momenta, p_r represents the radial velocity of the moving mass, p_θ denotes the modulus of the angular momentum vector, and p_ν is the polar component of the angular momentum. In addition to this, the symbol t stands for the physical time, and μ is the gravitational parameter of the central body.

Resorting to the polar nodal variables, and remembering the well known Keplerian orbital elements $(\, a\,,\, e\,,\, I\,,\, \omega\,,\, \Omega\,,\, M\,)$, the *Delaunay variables in hyperbolic motion* are

$$
\begin{aligned}
l &= e \sinh F - F = M\,, & L &= -\sqrt{\mu a}\,,\\
g &= \theta - f = \omega\,, & G &= \sqrt{\mu a (e^2 - 1)} = p_\theta\,,\\
h &= \nu = \Omega\,, & H &= G \cos I = p_\nu\,,
\end{aligned}
$$

with the following meaning for the subsidiary quantities $e(L, G)$ and $p(G)$, as functions of the Delaunay momenta, and the auxiliary variables $F(r; L, G)$, hyperbolic eccentric anomaly, and $f(r; L, G)$, true anomaly:

$$
\begin{aligned}
e^2 &= 1 + \frac{G^2}{L^2} > 1\,, & p &= \frac{G^2}{\mu} = a(e^2 - 1)\,,\\
r &= a(e \cosh F - 1)\,, & r &= \frac{p}{1 + e \cos f}\,.
\end{aligned}
$$

Notice that

$$\sqrt{e^2 - 1} = \frac{G}{(-L)},$$

with the positive determination of the square root on the left–hand side.

Other usual and helpful relations between the anomalies f and F are

$$\sin f = \frac{\sqrt{e^2 - 1}\,\sinh F}{e\cosh F - 1}, \qquad \cos f = \frac{e - \cosh F}{e\cosh F - 1},$$

$$\sinh F = \frac{\sqrt{e^2 - 1}\,\sin f}{1 + e\cos f}, \qquad \cosh F = \frac{e + \cos f}{1 + e\cos f},$$

and the (hyperbolic) Gauss equation

$$\tan\left(\frac{f}{2}\right) = \sqrt{\frac{e+1}{e-1}}\,\tanh\left(\frac{F}{2}\right).$$

In the next section the Delaunay canonical elements will be applied to the Hamiltonian of the *Main Problem*.

3. The Main Problem of Artificial Satellite Theory

Using the canonical set of *Hill–Whittaker polar nodal variables* $(r, \theta, \nu; p_r, p_\theta, p_\nu)$ to coordinatize the 6–dimensional phase space, the canonical formulation of the *Main Problem* of the theory of motion of *zonal satellites* leads to an investigation of the dynamical system governed by the Hamiltonian function

$$\begin{aligned}\mathcal{M} &= \mathcal{H}_0(r, -, -; p_r, p_\theta, -) + \varepsilon\,\mathcal{M}_1(r, \theta, -; -, p_\theta, p_\nu) \\ &= \frac{1}{2}\left[{p_r}^2 + \frac{{p_\theta}^2}{r^2}\right] - \frac{\mu}{r} + \varepsilon\,\frac{\mu\,{R_e}^2}{4r^3}\left\{(3c^2 - 1) + 3s^2\cos 2\theta\right\},\end{aligned}$$

with the customary abbreviations for the functions of the inclination $I \equiv I(p_\theta, p_\nu)$:

$$c \equiv c(p_\theta, p_\nu) = \cos I = \frac{p_\nu}{p_\theta}, \qquad s \equiv s(p_\theta, p_\nu) = \sin I.$$

As for the notations, the function $\mathcal{H}_0$ represents the Hamiltonian of a standard Kepler problem, R_e designates the mean equatorial radius of the central body, and the (small) parameter $\varepsilon = -J_2$ is a *dimensionless measure of the flattening of the primary*.

By virtue of the zonal nature of this problem, the polar component of the angular momentum, $p_\nu = H$, is an integral of the motion.

It is now remembered that, unlike the Delaunay variables, the applicability of the polar nodal variables is not restricted to the study of a specific kind of motion. Thus, we may contemplate the above Hamiltonian $\mathcal{M}$ as that corresponding to any type of orbit in

the context of the J_2 problem of the Artificial Satellite Theory. With all due precaution, and taking into account the different expressions of the relations defining the Delaunay variables and the Keplerian elements for elliptic and for hyperbolic orbital motion, we can formalize both cases of the Main Problem under a common Hamiltonian function, as shown below.

In terms of the Delaunay elements, the preceding Hamiltonian is formulated as

$$\begin{aligned}\mathcal{M} &= \mathcal{K}_0(-,-,-;L,-,-) + \varepsilon\,\mathcal{M}_1(l,g,-;L,G,H)\\ &= \mathcal{K}_0(L) + \varepsilon\,\frac{\mu\,{R_e}^2}{4r^3}\left\{(3c^2-1) + 3s^2\cos 2\theta\right\},\end{aligned}$$

where $\mathcal{K}_0$ is the Keplerian Hamiltonian in Delaunay variables, while r and $\theta = f + g$ are understood to be expressed in terms of the appropriate Delaunay set, taking into account the form of the Keplerian orbital elements as subsidiary quantities and the auxiliary variables f and F (or f and E, in case of elliptic motion).

For convenience, and in anticipation of future calculations, one usually rewrites the first–order part of $\mathcal{M}$ under the form

$$\begin{aligned}\mathcal{M}_1 &= \varepsilon\,\frac{\mu\,{R_e}^2}{4a^3}\left\{(3c^2-1)\left(\frac{a}{r}\right)^3 + 3s^2\left(\frac{a}{r}\right)^3\cos 2\theta\right\}\\ &= \varepsilon\,\frac{\mu^4\,{R_e}^2}{4L^6}\left\{(3c^2-1)\left(\frac{a}{r}\right)^3 + 3s^2\left(\frac{a}{r}\right)^3\cos 2\theta\right\}.\end{aligned}$$

4. Elimination of the mean anomaly

In what follows, we return to the consideration of the hyperbolic J_2 problem. Resorting to an appropriate modification of the *Lie transform technique* (Hori, 1966; Deprit, 1969), the outlines of *a first–order analytical integration based on the adequate Delaunay variables* follow the pattern presented by Hori (1961).

Correspondingly, in order to perform a first–order Delaunay normalization of the positive energy Main Problem, a canonical transformation to a new set of variables of the same type as those in which the problem is formulated, say

$$(l,g,h;L,G,H) \quad \xrightarrow{W} \quad (l',g',h';L',G',H),$$

is proposed that is governed by a generating function

$$W \equiv W(l',g',h';L',G',H) = \varepsilon\,W_1(l',g',h';L',G',H) + \mathcal{O}(\varepsilon^2),$$

whose specification is effected by a modification of the Lie transform method. For the purposes of the present study, the determination of W up to the first order will suffice.

(Remember also that H is an integral of the problem, and so it remains unchanged under this transformation).

It is desired that the transformed Hamiltonian should take on the form

$$\mathcal{M} \xrightarrow{W} \mathcal{M}' = \mathcal{M}_0{}' = \mathcal{K}_0{}' = \mathcal{K}_0(L') = \frac{\mu^2}{2(L')^2} .$$

The primed quantities are obtained from the original ones by putting the corresponding new (primed) canonical variable in place of the old one. For the sake of conciseness in the notation, the prime symbol will be dropped out. Thus, the first–order equation of the perturbation method will now read

$$\{\mathcal{K}_0, W_1\} = -\mathcal{M}_1 \Longrightarrow -\frac{\partial \mathcal{K}_0}{\partial L}\frac{\partial W_1}{\partial l} = -\mathcal{M}_1(l, g, -; L, G, H)$$

$$\Longrightarrow \frac{\partial W_1}{\partial l} = -\frac{\mu^2 {R_e}^2}{4L^3}\left[A\left(\frac{a}{r}\right)^3 + B\left(\frac{a}{r}\right)^3 \cos(2g + 2f)\right],$$

bearing in mind that the quantities occurring in these expressions depend on the new variables. In particular, one has introduced the notations

$$A = 3c^2 - 1 = 3\frac{H^2}{G'^2} - 1, \quad B = 3s^2 = 3\left(1 - c^2\right) = 3\left(1 - \frac{H^2}{G'^2}\right).$$

The reckoning work can be carried out essentially in the same way as in Hori's 1961 paper, taking into account that now we are dealing with the new (primed) variables instead of considering the old Delaunay angles and the new momenta.

The final specification of W_1 and the perturbation study can be carried out as in Hori's article, which requires the knowledge of certain partial derivatives. The detail of some steps in the calculation process seems to be rather elusive. In this respect, we have never found any clarifying remark or hint in the literature. This is the reason why we intend to elucidate the dependence of the functions involved in the calculation.

In a more precise way, partial derivatives with respect to the angles and with respect to H pose no problem. On the other hand, by application of the chain rule, the partial derivatives with respect to the new canonical momenta L' and G' (see Brouwer 1959, pp. 378–379, and Hori 1961, p.260) are reduced to partial derivatives with respect to the eccentricity–like quantity e (read e'). This is just the point that we wish to clarify, since we think that the inconvenience we have encountered throughout the computation of the said derivatives is mainly due to the form under which certain functional dependences are nested.

5. On Certain Functional Dependences in the Kepler Problem

The analytical treatment of some perturbed Keplerian systems (e. g., the planetary theory, the theory of motion of an artificial satellite), say the way of developing approximate analytical solutions when dealing with perturbed Keplerian systems, usually resorts to methods involving a Delaunay normalization.

With this aim in view, some mathematical tools are required. Some authors (Ahmed, 1994; Kelly, 1989) have recently contributed formulae and techniques to deal with the reduction of perturbed Keplerian systems to normal form. The said authors have evaluated certain integrals occurring when performing this process of reduction. In his turn, Palacián (1992), Appendix 2, gives a table of derivatives with respect to the Hill–Whittaker and Delaunay variables.

In performing a Delaunay normalization of the artificial satellite problem, for elliptic motion Brouwer (1959), p. 379, states that

$$\frac{\partial}{\partial e}\left(\frac{a}{r}\right) = \left(\frac{a}{r}\right)^2 \cos f\,, \quad \frac{\partial f}{\partial e} = \left(\frac{a}{r} + \frac{L^2}{G^2}\right) \sin f\,,$$

while for hyperbolic motion Hori (1961), p. 260, adduces

$$\frac{\partial}{\partial e}\left(\frac{a}{r}\right) = -\left(\frac{a}{r}\right)^2 \cos f\,, \quad \frac{\partial f}{\partial e} = -\left(\frac{a}{r} + \frac{L^2}{G^2}\right) \sin f\,.$$

We have encountered certain difficulties when trying to deduce these and other formulae, whose calculation seems to be rather elusive. In order to reconstruct these derivatives for application to hyperbolic–like motion (the treatment of the elliptic case is analogous), we establish the functional dependences through which we shall interpret and carry out the calculation.

We start from the basic formulae

$$a \equiv a(L) = L^2/\mu\,, \quad p \equiv p(G) = G^2/\mu\,,$$

$$e \equiv e(L, G) = \sqrt{1 + \frac{G^2}{L^2}}\,, \quad e \equiv e(a, p) = \sqrt{1 + \frac{p}{a}}\,,$$

$$p = a(e^2 - 1) \implies \frac{p}{a}(L, G) = e^2 - 1 = \frac{G^2}{L^2}\,.$$

For convenience, we introduce

$$r \equiv r(l; L, G) = \frac{p}{1 + e\cos f} \iff \frac{p}{r} \equiv \frac{p}{r}(e, l) = 1 + e\cos f\,,$$

$$r \equiv r(l; L, G) = a(e\cosh F - 1) \iff \frac{a}{r} \equiv \frac{a}{r}(e, l) = \frac{1}{e\cosh F - 1}\,.$$

A significant role is played by the *(hyperbolic) Kepler equation*

$$l = e \sinh F - F .$$

Observe that we are considering $f \equiv f(e,l)$ and $F \equiv F(e,l)$ through the above Kepler equation and the customary relations between f and F.

With these conventions, in the next section we undertake the construction of the desired derivatives.

6. Calculation of Some Partial Derivatives

Successive steps will complete a set of formulae which will be applied in future perturbation developments.

6.1 Calculation of $\partial F/\partial e$

According to the preceding remark concerning the dependence of $F \equiv F(e,l)$ through the Kepler equation, by forming the partial derivative with respect to e in that equation we get

$$0 = \frac{\partial l}{\partial e} = \sinh F + (e \cosh F - 1)\frac{\partial F}{\partial e} ,$$

from which there results

$$\frac{\partial F}{\partial e} = \frac{-\sinh F}{(e \cosh F - 1)} = -\frac{a}{r}\sinh F .$$

6.2 Calculation of $\partial(a/r)/\partial e$ and $\partial(r/a)/\partial e$

From $r = a(e \cosh F - 1)$ we deduce

$$\frac{\partial}{\partial e}\left(\frac{a}{r}\right) = \frac{\partial}{\partial e}\left(\frac{1}{e \cosh F - 1}\right) = -\left(\frac{a}{r}\right)^2\left(\cosh F + e \sinh F \frac{\partial F}{\partial e}\right)$$

$$= -\left(\frac{a}{r}\right)^2\left(\cosh F - \frac{e \sinh^2 F}{(e \cosh F - 1)}\right) = -\left(\frac{a}{r}\right)^2 \frac{e - \cosh F}{(e \cosh F - 1)} .$$

And finally

$$\frac{\partial}{\partial e}\left(\frac{a}{r}\right) = -\left(\frac{a}{r}\right)^2 \cos f .$$

On the other hand, since

$$\frac{r}{a}\frac{a}{r} = 1 \implies \frac{a}{r}\frac{\partial}{\partial e}\left(\frac{r}{a}\right) + \frac{r}{a}\frac{\partial}{\partial e}\left(\frac{a}{r}\right) = 0 ,$$

by solving for $\partial(r/a)/\partial e$ we conclude that

$$\frac{\partial}{\partial e}\left(\frac{r}{a}\right) = -\left(\frac{r}{a}\right)^2 \frac{\partial}{\partial e}\left(\frac{a}{r}\right) = \cos f\,.$$

This formula may also be obtained in the following simple way: starting from

$$\frac{\partial}{\partial e}\left(\frac{r}{a}\right) = \frac{\partial}{\partial e}\left(e\cosh F - 1\right) = \cosh F + e\sinh F\,\frac{\partial F}{\partial e}\,,$$

by applying the preceding result (Subsection 6.1) concerning the form of $\partial F/\partial e$ and the Keplerian relations between the anomalies, we arrive at the desired final expression.

6.3 Calculation of $\partial f/\partial e$

Remember that we are considering $f \equiv f(e,l)$ and $F \equiv F(e,l)$, and the basic relation

$$\cos f = \frac{e - \cosh F}{e\cosh F - 1}\,.$$

By constructing here the partial derivative of both sides of this equality with respect to e, one has

$$-\sin f\,\frac{\partial f}{\partial e} = \frac{\left(1 - \sinh F\,(\partial F/\partial e)\right)\left(e\cosh F - 1\right)}{\left(e\cosh F - 1\right)^2} - \frac{\left(\cosh F + e\sinh F\,(\partial F/\partial e)\right)\left(e - \cosh F\right)}{\left(e\cosh F - 1\right)^2}\,;$$

after introducing the expression previously obtained for $\partial F/\partial e$, we eliminate this derivative and get

$$-\sin f\,\frac{\partial f}{\partial e} = \frac{a}{r} + \frac{\sin^2 f}{e^2-1} - \frac{a}{r}\left(\frac{e - \cosh F}{e\cosh F - 1}\right)^2$$

$$= \frac{a}{r} + \frac{\sin^2 f}{e^2-1} - \frac{a}{r}\cos^2 f = \left(\frac{1}{e^2-1} + \frac{a}{r}\right)\sin^2 f\,,$$

and finally

$$\frac{\partial f}{\partial e} = -\left(\frac{1}{e^2-1} + \frac{a}{r}\right)\sin f\,.$$

This expression can also be recovered by performing the calculations in a different way, which suggests an alternate derivation for this formula: starting from

$$\tan\frac{f}{2} = \sqrt{\frac{e+1}{e-1}}\,\tanh\frac{F}{2}\,,$$

and taking the logarithmic derivative with respect to e, one obtains

$$\frac{1}{2\tan(f/2)\cos^2(f/2)}\,\frac{\partial f}{\partial e} = -\frac{1}{e^2-1} + \frac{1}{2\tanh(F/2)\cosh^2(F/2)}\,\frac{\partial F}{\partial e}\,.$$

Consequently, the double–angle formulae yield

$$\frac{1}{\sin f}\,\frac{\partial f}{\partial e} = -\,\frac{1}{e^2-1} + \frac{1}{\sinh F}\,\frac{\partial F}{\partial e},$$

and keeping in mind (Subsection 6.1) the expression for $\partial F/\partial e$:

$$\frac{\partial f}{\partial e} = -\left(\frac{1}{e^2-1} + \frac{a}{r}\right)\sin f\,.$$

6.4 Calculation of $\partial(p/r)/\partial e$ and $\partial(r/p)/\partial e$

Taking into account the relation

$$p = a(e^2-1) \implies \frac{r}{a}\,\frac{p}{r} = e^2-1,$$

the partial derivative with respect to e gives

$$\frac{r}{a}\,\frac{\partial}{\partial e}\left(\frac{p}{r}\right) + \frac{p}{r}\,\frac{\partial}{\partial e}\left(\frac{r}{a}\right) = 2e.$$

By solving for the sought derivative

$$\frac{\partial}{\partial e}\left(\frac{p}{r}\right) = \frac{a}{r}\left[2e - \frac{p}{r}\,\frac{\partial}{\partial e}\left(\frac{r}{a}\right)\right] = \frac{a}{r}\left(2e - \frac{p}{r}\cos f\right).$$

For an alternative derivation, we consider that $a/p = 1/(e^2-1)$ can be looked on as a function depending on e only, and then

$$\frac{d}{de}\left(\frac{a}{p}\right) = -\,2e\left(\frac{a}{p}\right)^2,$$

from which we have

$$\frac{\partial}{\partial e}\left(\frac{r}{p}\right) = \frac{\partial}{\partial e}\left(\frac{r}{a}\,\frac{a}{p}\right) = \frac{r}{a}\,\frac{d}{de}\left(\frac{a}{p}\right) + \frac{a}{p}\,\frac{\partial}{\partial e}\left(\frac{r}{a}\right),$$

and so

$$\frac{\partial}{\partial e}\left(\frac{r}{p}\right) = \frac{1}{e^2-1}\left(-2e\,\frac{r}{p} + \cos f\right) = \left(\frac{L}{G}\right)^2\left(-2e\,\frac{r}{p} + \cos f\right).$$

Notice also that

$$\frac{p}{r}\,\frac{r}{p} = 1 \implies \frac{\partial}{\partial e}\left(\frac{p}{r}\right) = -\left(\frac{p}{r}\right)^2\frac{\partial}{\partial e}\left(\frac{r}{p}\right) = \frac{1}{e^2-1}\,\frac{p}{r}\left(2e - \frac{p}{r}\cos f\right).$$

7. Final Summary of Formulae

With the help of the auxiliary relations

$$\sqrt{e^2-1}=\frac{G}{(-L)}\,,\qquad \frac{a}{r}=\left(\frac{L}{G}\right)^2\frac{p}{r};$$

$$\sin f=\frac{\sqrt{e^2-1}\,\sinh F}{e\cosh F-1}\quad\Longrightarrow\quad \sin f=-\frac{G}{L}\frac{a}{r}\sinh F=-\frac{L}{G}\frac{p}{r}\sinh F\,,$$

we complete the set of formulae:

$$\frac{\partial F}{\partial e}=-\frac{\sinh F}{e\cosh F-1}=-\frac{a}{r}\sinh F=\frac{L}{G}\sin f=-\left(\frac{L}{G}\right)^2\frac{p}{r}\sinh F\,,$$

$$\begin{aligned}\frac{\partial f}{\partial e}&=-\left(\frac{1}{e^2-1}+\frac{a}{r}\right)\sin f=-\left(\frac{L^2}{G^2}+\frac{a}{r}\right)\sin f=-\frac{L^2}{G^2}\left(\frac{p}{r}+1\right)\sin f\\&=\frac{G}{L}\frac{a}{r}\left[\left(\frac{a}{r}\right)+\left(\frac{L}{G}\right)^2\right]\sinh F=\left(\frac{L}{G}\right)^3\frac{p}{r}\left[\left(\frac{p}{r}\right)+1\right]\sinh F\,,\end{aligned}$$

$$\frac{\partial}{\partial e}\left(\frac{p}{r}\right)=\frac{a}{r}\left(2e-\frac{p}{r}\cos f\right)=\frac{a}{r}\left[2e-\left(\frac{G}{L}\right)^2\frac{a}{r}\cos f\right]=\left(\frac{L}{G}\right)^2\frac{p}{r}\left(2e-\frac{p}{r}\cos f\right).$$

Acknowledgements

The authors thank their colleague Dr. Pablo Martín for his advice in the preparation of final version of the manuscript. Partial financial support for the second author (Floría) came from the National Programme of Space Research of CICYT, Comisión Interministerial de Ciencia y Tecnología of Spain, under Project ESP 93–741.

Authors are listed in alphabetical order.

References

[1] Brouwer, D.: 1959, Solution of the Problem of Artificial Satellite Theory without Drag, *Astronomical Journal* **64,** pp. 378–397.

[2] Brouwer, D. and Clemence, G. M.: 1961, *Methods of Celestial Mechanics*, Academic Press, New York and London.

[3] Deprit, A.: 1969, Canonical Transformations Depending on a Small Parameter, *Celest. Mech.* **1,** pp. 12–30.

[4] Deprit, A.: 1981, The Elimination of the Parallax in Satellite Theory, *Celest. Mech.* **24,** pp. 111–153.

[5] Deprit, A.: 1982, Delaunay Normalisations, *Celest. Mech.* **26,** pp. 9–21.

[6] Floría, L.: 1990, Variables de tipo DS para órbitas hiperbólicas. In: *Actas das XV Jornadas Luso–Espanholas de Matemática (Evora, Portugal, September 1990)*, vol. **V,** pp. 251–256. (Spanish language; abstract in English. Available from the author).

[7] Floría, L.: 1993, *Intermediarios radiales y generalizaciones de las variables de tipo Delaunay–Scheifele. Aplicación al movimiento orbital de satélites artificiales* (Ph. D. Thesis). Universidad de Valladolid (Spain).

[8] Hori, G.–i.: 1961, The Motion of a Hyperbolic Artificial Satellite around the Oblate Earth, *Astronomical Journal* **66,** pp. 258–263.

[9] Hori, G.–i.: 1966, Theory of General Perturbations with Unspecified Canonical Variables, *Publications of the Astronomical Society of Japan* **18,** pp. 287–296.

[10] Palacián, J.: 1992, *Teoría del satélite artificial: armónicos teserales y su relegación mediante simplificaciones algebraicas* (Ph. D. Thesis). Universidad de Zaragoza (Spain).

Rev. Academia de Ciencias. Zaragoza. 49 (1994)

Respuesta aleatorizada en muestreo estratificado y para estudios analíticos.

M. Ruiz Espejo y A. Arcos Cebrián

Deptº de Estadística e Investigación Operativa
Facultad de Ciencias Económicas y Empresariales
Universidad Complutense. 28223 - MADRID

Deptº de Estadística e Investigación Operativa
Facultad de Ciencias. Universidad de Granada
18071 - GRANADA.

Summary.

We study the method of randomized response for analytic studies proposing an unbiased estimator for affirmative proportion of a certain intimate question, and calculating its variance. An analogously study is carried out in stratified sampling, obtaining the minimum allocation in this context. We conclude with a study of randomized response combining, simultaneously, stratified sampling and analytic studies.

1. Introducción.

El problema de respuesta aleatorizada, propuesto por Warner (1965), no ha sido estudiado cuando nos interesamos en estudios analíticos (Koop, 1986), ni se ha tratado en combinación con muestreo estratificado, o con muestreo estratificado para estudios analíticos (Ruiz, 1991). En este trabajo presentamos, para los tres casos anteriores, estimadores insesgados de la proporción de la población infinita subyacente que contestaría afirmativamente a una cuestión de carácter íntimo. Asimismo, calcularemos la varianza de tales estimadores.

2. Respuesta aleatorizada para estudios analíticos.

Supongamos que las posibles respuestas de una población finita ("si" o "no"), constituyen una materialización de una superpoblación, de

modo que la población finita, de tamaño N, se puede considerar como una muestra aleatoria simple de una superpoblación infinita que tiene probabilidad A ó C de responder "si" a la pregunta íntima o a la pregunta aleatorizada, respectivamente. Si B es la respuesta afirmativa conocida de cierta pregunta intrascendente en la población finita observable y P la probabilidad de preguntar la cuestión íntima a un individuo (siendo 1-P la probabilidad de preguntar la cuestión intrascendente al individuo), resulta que la proporción de respuestas afirmativas con la pregunta aleatorizada en la población finita es, por el teorema de la probabilidad total,

$$C = PA + (1-P)B \qquad (1)$$

siendo A y B las proporciones, en la población finita, de "síes" en la cuestión íntima e intrascendente, respectivamente. Despejando A en (1), tenemos

$$A = \frac{C}{P} - \frac{(1-P)B}{P}$$

El estimador natural insesgado para A, será:

$$\hat{A} = \frac{\hat{C}}{P} - \frac{(1-P)B}{P}$$

Usando diseño "masr" (muestreo aleatorio simple con reemplazamiento) y siendo M el modelo superpoblacional latente en estudios analíticos (Koop, 1986), tenemos

$$E(\hat{A}) = E_M E_{masr}(\hat{A}) = E_M(A) = \bar{A}$$

es decir, A es insesgado para el parámetro superpoblacional A. Para calcular la varianza de A, aplicamos el teorema de Madow,

$$V(\hat{A}) = E_M V_{masr}(\hat{A}) + V_M E_{masr}(\hat{A}) \qquad (2)$$

$$V_{masr}(\hat{A}) = \frac{1}{P^2}\,\frac{C(1-C)}{n}$$

de donde

$$E_M V_{masr}(\hat{A}) = \frac{1}{nP^2} E_M(C-C^2) = \frac{1}{nP^2}\{\bar{C}-[\bar{C}^2+V_M(C)]\} =$$

$$= \frac{1}{nP^2}\left(\bar{C}-\bar{C}^2 - \frac{\bar{C}(1-\bar{C})}{N}\right) = \frac{1}{nP^2}(\bar{C}-\bar{C}^2)\,\frac{N-1}{N} \qquad (3)$$

Por otra parte, se tiene

$$E_{masr}(\hat{\bar{A}}) = \bar{A} \qquad y \qquad V_M E_{masr}(\hat{\bar{A}}) = V_M(\bar{A}) = \frac{\bar{A}(1-\bar{A})}{N} \tag{4}$$

Sustituyendo las igualdades (3) y (4) en (2), resulta

$$V(\hat{\bar{A}}) = \frac{N-1}{nNP^2}(\bar{C}-\bar{C}^2) + \frac{\bar{A}-\bar{A}^2}{N} \tag{5}$$

donde $\bar{C}-\bar{C}^2 = NV_M(C)$ y $\bar{A}-\bar{A}^2 = NV_M(A)$. La expresión (5) es la varianza de $\hat{\bar{A}}$ como estimador insesgado de $\bar{A}$ en estudios analíticos, con pregunta aleatorizada en la primera fase.

3. Respuesta aleatorizada en muestreo estratificado.

En el estrato h (de tamaño N_h) de una población finita, tenemos

$$C_h = P_h A_h + (1-P_h)B_h$$

siendo P_h la probabilidad de preguntar la cuestión íntima y A_h, B_h, C_h, las proporciones de contestación afirmativa en el estrato h de la pregunta íntima, intrascendente y aleatorizada, respectivamente. Entonces, el estimador insesgado de C_h será su proporción muestral $\hat{C}_h$ (La pregunta intrascendente cuya proporción de "síes" en el estrato h, es decir B_h, puede ser "una" común, ó "varias" de un estrato a otro). Ahora tendremos

$$\hat{C}_h = P_h \hat{A}_h + (1-P_h)B_h$$

de donde un estimador insesgado de A_h, será

$$\hat{A}_h = \frac{1}{P_h}[\hat{C}_h - (1-P_h)B_h]$$

siendo su varianza

$$V(\hat{A}_h) = \frac{1}{P_h^2} V(\hat{C}_h) = \frac{1}{P_h^2} \frac{C_h(1-C_h)}{n_h}$$

y donde n_h es el tamaño muestral con diseño "masr" en el estrato h.

El estimador insesgado de A, la proporción de "síes" de la cuestión íntima en la población finita, será

$$\hat{A} = \sum_{h=1}^{L} W_h \hat{A}_h$$

donde $W_h = N_h/N$ (h = 1,2,...,L) y L el número de estratos. La varianza es

$$V(\hat{A}) = \sum_{h=1}^{L} W_h^2 V(\hat{A}_h) = \sum_{h=1}^{L} W_h^2 \frac{1}{P_h^2} \frac{C_h(1-C_h)}{n_h}$$

y puede minimizarse por el método de los multiplicadores de Lagrange. Así, llamando n al tamaño muestral global, esto es

$$n = \sum_{h=1}^{L} n_h$$

se obtiene

$$n_h \propto \frac{W_h \sqrt{C_h(1-C_h)}}{P_h} \qquad (h = 1,2,...,L)$$

de donde la varianza mínima es

$$V_{mín}(\hat{A}) = \frac{1}{n} \left(\sum_{h=1}^{L} \frac{W_h \sqrt{C_h(1-C_h)}}{P_h} \right)^2$$

Si $P_h = 1$ para todo h, se pregunta con seguridad la cuestión íntima. Si $0 < P_h < 1$ para algún h, aumenta la protección de la intimidad del encuestado, pero se pierde precisión del estimador A, que puede recuperarse aumentando suficientemente el tamaño muestral n.

4. Respuesta aleatorizada en muestreo estratificado para estudios analíticos.

El estimador propuesto en este caso, será

$$\hat{\bar{A}} = \sum_{h=1}^{L} \bar{W}_h \hat{A}_h$$

donde $\bar{W}_h$ es el peso relativo superpoblacional en el estrato h, y $\hat{A}_h$ la proporción poblacional estimada que contesta afirmativamente a la

pregunta íntima en el estrato h. Sea N_h el tamaño del estrato h en la población finita, de modo que si L es el número de estratos, tendremos

$$N = \sum_{h=1}^{L} N_h$$

El estimador $\hat{\bar{A}}$ es insesgado para $\bar{A}$ cuando, en cada estrato, se muestrea independiente y aleatoriamente con reposición. En efecto,

$$E(\hat{\bar{A}}) = \sum_{h=1}^{L} \bar{W}_h E_M E_{masr}(\hat{\bar{A}}_h) = \sum_{h=1}^{L} \bar{W}_h \bar{A}_h = \bar{A}$$

siendo $\bar{A}_h$ la probabilidad de respuesta afirmativa a la pregunta íntima en el estrato superpoblacional h-ésimo. Utilizando la igualdad (5), en cada estrato la varianza de $\hat{\bar{A}}$ en estudios analíticos, es

$$V(\hat{\bar{A}}) = \sum_{h=1}^{L} \bar{W}_h V(\hat{\bar{A}}_h) = \sum_{h=1}^{L} W_h^2 \left(\frac{(N_h-1)(\bar{C}_h-\bar{C}_h^2)}{n_h N_h P_h^2} + \frac{\bar{A}_h-\bar{A}_h^2}{N_h} \right) \qquad (6)$$

siendo n_h el tamaño muestral en el estrato h, $\bar{C}_h$ la probabilidad de respuesta afirmativa a la pregunta aleatorizada en el estrato superpoblacional h-ésimo, y P_h la probabilidad de pregunta íntima en el estrato h.

Minimizando la varianza obtenida en (6), sujeto a que

$$n = \sum_{h=1}^{L} n_h$$

deducimos (aplicando el método de los multiplicadores de Lagrange) que debe guardarse la siguiente proporcionalidad

$$n_h \propto \bar{W}_h \sqrt{(N-1)(\bar{C}_h-\bar{C}_h^2)/P_h \sqrt{N_h}} \qquad (7)$$

Finalmente, sustituyendo (7) en (6), la varianza mínima es

$$V_{mín}(\hat{\bar{A}}) = \frac{1}{n} \left(\sum_{h=1}^{L} \bar{W}_h \sqrt{(N-1)(\bar{C}_h-\bar{C}_h^2)/P_h \sqrt{N_h}} \right)^2 + \sum_{h=1}^{L} \bar{W}_h^2 \frac{\bar{A}_h-\bar{A}_h^2}{N_h} \qquad (8)$$

Observar que, en cada estrato, B_h (proporción de respuestas afirmativas a la pregunta intrascendente en el estrato h de la población finita) debe ser conocida, pudiendo variar la pregunta de un estrato a otro.

Si $0 < P_h < 1$ para algún h, la varianza dada por (8) es mayor respecto del caso $P_h = 1$. De cualquier modo, si $P_h > 0$ para todo h, al aumentar el tamaño muestral n, la precisión puede recuperarse a un nivel mayor de protección de la intimidad de los encuestados, pues el segundo monomio en (8) es constante.

Referencias

[1] Koop, J.C. (1986). Some problems of statistical inference from sample survey data for analytic studies. *Statistics* **17**, 237-247.

[2] Ruiz, M. (1991). Muestreo estratificado en estudios analíticos. *Metron* **49**, 459-468.

[3] Ruiz, M. y Ruiz, M.M. (1991). Una nota sobre la protección de la intimidad con respuesta aleatorizada. *Qüestiió* **15**, 47-53.

[4] Warner, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.* **60**, 63-69.

Rev. Academia de Ciencias. Zaragoza. 49 (1994)

Sobre la invarianza lineal en problemas de estratificación óptima

M. Ruiz Espejo y A. Arcos Cebrián

Dept° de Estadistica e Investigación Operativa
Facultad de Ciencias Económicas y Empresariales
Universidad Complutense. 28223 - MADRID

y

Dept° de Estadistica e Investigación Operativa
Facultad de Ciencias. Universidad de Granada
18071 - GRANADA

Summary.

The linear invariance of optimum stratification bounds are studied for estimating population mean and variance; moreover, the invariance of optimum sample allocations and the proportionality of usual variance estimators in each stratum, allow us of the stratification techniques of infinite population, to have valid instruments for a population and all transformed from this to origin and scale changes in the measure of the units, that is to say, the techniques are communly valid for wide families of populations.

1. Introducción.

Hasta ahora se conocen dos problemas de estratificación óptima, planteados para minimizar la varianza de estimadores insesgados en muestreo estratificado con el objetivo de estimar la media poblacional y la varianza poblacional, respectivamente. El problema de estratificación óptima para estimar la media poblacional, es debido a Dalenius, T. (1957, cap. 7), mientras que para estimar la varianza poblacional, es debido a Ruiz, M y Ruiz, M.M. (1992). En ambos casos se producen sustanciales ganancias de precisión con respecto a los estimadores media y cuasivarianza muestrales clásicos.

Su aplicación actual en problemas de inferencia estadística es relativa, pues a pesar de dar pautas para minimizar la varianza de los estimadores insesgados habituales, su uso práctico sólo se hace mediante aproximaciones, que si bien dejan ya de dar soluciones óptimas, producen ganancias suficientes en precisión como para considerarlas útiles en problemas interferenciales concretos.

En orden a hacer buen uso de estas técnicas, vamos a justificar que los puntos de estratificación óptima son invariantes lineales y que, las afijaciones muestrales y los tamaños relativos de los estratos no varían al hacer transformaciones lineales en la variable aleatoria considerada; por último, las varianzas de los estimadores insesgados para poblaciones transformadas linealmente son directamente proporcionales a la varianza de una población fija, previamente considerada. De este modo, si disponemos de una población fija, las soluciones que obtendremos con ella serán análogas a las que resulten para cualquier otra población transformada linealmente de la anterior. Estas ideas nos permiten dar los pasos a realizar ante un problema concreto de inferencia o estimación estratificada de la media o varianza poblacional cuando se conoce el tipo de distribución poblacional, salvo un cambio de origen y escala.

2. Estimación de la media poblacional.

Partimos de la variable aleatoria Z con función de densidad $f(z)$, $z \in \mathbf{R}$. Esta variable aleatoria es $Z = a + bY$ (a real y b positiva), siendo Y la variable aleatoria de interés. Entonces, la función de densidad $g(y)$ (y real) de la variable aleatoria Y será para todo y real

$$g(y) = bf(a+by) \qquad (2.1)$$

Para verlo, sea F la función de distribución de la variable aleatoria Z,

$$F(z) = p(Z \leq z) = p\left(\frac{Z-a}{b} = Y \leq \frac{z-a}{b}\right) = G\left(\frac{Z-a}{b}\right)$$

por lo que

$$f(z) = F'(z) = G'\left(\frac{z-a}{b}\right)\frac{1}{b} = G'(y)\frac{1}{b} = g(y)\frac{1}{b}$$

siendo G la función de distribución de la variable aleatoria Y.

2.1. Invarianza de los pesos relativos.

Sea $W_h(z)$ el peso o tamaño relativo del estrato h , para la variable aleatoria Z ; y sea R_h el recorrido del estrato h en la variable aleatoria Z. Haciendo el cambio $Z = a+bY$, el nuevo recorrido para la variable aleatoria Y , será R'_h , y de acuerdo con (2.1), tendremos

$$W_h(z) = \int_{R_h} f(z)dz = \int_{R'_h} f(a+by)bdy = \int_{R'_h} g(y)dy = W_h(y) \qquad (2.2)$$

luego los pesos relativos no varían para cambios lineales con b positivo.

2.2. Invarianza de la afijación mínima.

Veamos primero que para las varianzas de los estratos se verifica la igualdad

$$\sigma_h^2(z) = b^2 \sigma_h^2(y)$$

En efecto, según (2.2), tendremos

$$\sigma_h^2(z) = \frac{1}{W_h(z)} \int_{R_h} [z-\mu_h(y)]^2 f(z)dz =$$

$$= \frac{1}{W_h(y)} \int_{R'_h} b^2[y-\mu_h(y)]^2 g(y)dy = b^2 \sigma_h^2(y)$$

ya que las medias de los estratos verifican la relación $\mu_h(z) = a + b\mu_h(y)$.

Como consecuencia (para L estratos), para todo $h = 1,2,...,L$:

$$n_h(z) = n \frac{W_h(z)\sigma_h(z)}{\sum_{k=1}^{L} W_k(z)\sigma_k(z)} = n \frac{W_h(y)b\sigma_h(y)}{\sum_{k=1}^{L} W_k(z)b\sigma_k(y)} = n_h(y)$$

2.3. Proporcionalidad de la varianza del estimador.

Siendo $\bar{z}'$ e $\bar{y}'$ los estimadores insesgados usuales en muestreo estratificado, sus varianzas con afijación mínima verifican la relación

$$V_{mín}(\bar{z}') = \frac{1}{n}\left(\sum_{h=1}^{L} W_h(z)\sigma_h(z)\right)^2 = \frac{1}{n}\left(\sum_{h=1}^{L} W_h(y)b\sigma_h(y)\right)^2 = b^2 V_{mín}(\bar{y}')$$

y como consecuencia, estas varianzas se minimizan para los correspondientes puntos de estratificación que pueden obtenerse directamente mediante el cambio lineal $z = a + by$. Es decir que, los puntos de estratificación son invariantes lineales, y la varianza mínima es directamente proporcional a cualquier varianza mínima del estimador para la variable aleatoria relacionada de forma lineal, o sea, para amplias familias de poblaciones. Es importante observar que la regla "cum $\sqrt{f}$" (ver Särndal, C.E. et al, 1991, p. 463), que aproxima los puntos de estratificación óptima para la función de densidad f , es también invariante ante cambios lineales de la variable aleatoria.

3. Estimación de la varianza poblacional.

Siguiendo la notación de Ruiz, M y Ruiz, M.M. (1992), ante un cambio lineal del tipo $z = a + by$ (a real y b positivo), el momento central de orden 4 en el estrato h, verifica las igualdades

$$\tau_{h,4}(z) = b^4 \tau_{h,4}(y) \qquad \sigma_h^4(z) = b^4 \sigma_h^4(y)$$

Por tanto, para $h = 1,2,...,L$, $n_h(z) = n_h(y)$, y la afijación óptima dada por Ruiz, M. y Ruiz, M.M. (1992, secc. 4) es invariante, así como los pesos relativos de los estratos: $W_h(z) = W_h(y)$. Consecuentemente, coinciden los mínimos de las varianzas siguientes:

$$V_{\text{ópt}}\left(\hat{\sigma}_{st}^2(z)\right) = b^4 V_{\text{ópt}}\left(\hat{\sigma}_{st}^2(y)\right)$$

ya que son directamente proporcionales sus sumando respectivos, con idéntica constante de proporcionalidad b^4.

La invarianza lineal de los puntos de estratificación equilibrada óptima, usados para estimar la varianza poblacional, como se sugirió en Ruiz, M. y Ruiz, M.M. (1992), es consecuencia inmediata del teorema 4.1 de Ruiz, M. (1990).

4. Conclusión.

Las técnicas de estratificación óptima son básicamente invariantes lineales, de manera que, aún desconociendo la función de densidad exacta "g" de un problema concreto, pero si la función de densidad "f" de un cambio lineal de ésta, podemos trabajar con la función "f", determinando las características principales de su estratificación óptima, y estimando las constantes "a" y "b" del cambio

lineal. De esta forma tendremos los elementos necesarios para su aplicación a la población caracterizada por la función "g".

La estimación de las constantes "a" y "b" puede realizarse del siguiente modo:

Para estimar $\mu(y)$, podemos partir de una muestra piloto aleatoria simple (muestra piloto reutilizable en el proceso inferencial con estratificación óptima), de manera que para inferir sobre $\mu(y)$ y $\sigma(y)$, bastaría exigir que

$$\mu(z) = a + b\hat{\mu}(y) \qquad y \qquad \sigma(z) = b\hat{\sigma}(y) \tag{4.1}$$

de donde

$$b = \frac{\sigma(z)}{\hat{\sigma}(y)} \qquad y \qquad \hat{a} = \mu(z) - b\hat{\mu}(y)$$

Entonces, para estimar $\sigma^2(y)$, siendo $\mu(y)$ conocida, tendremos

$$\hat{\sigma}^2(y) = \frac{1}{n}\sum_{i=1}^{n}[y_i^{(0)} - \mu(y)]^2$$

siendo $y_i^{(0)}$ la i-ésima observación de la muestra aleatoria simple piloto.

Los pasos a realizar para completar la inferencia, son:

1. Proponer la forma de la función de densidad f(z) (z real).

2. Determinar los puntos de estratificación óptima para la variable aleatoria Z, conocida.

3. Calcular los tamaños relativos $W_h(z)$, h = 1,2,...,L, de los estratos así construidos, como la afijación óptima $n_h(z)$, h = 1,2,...,L, ya sea para estimar μ ó σ^2.

4. Obtener una muestra piloto aleatoria simple de la variable aleatoria de interés Y, estimando "a" y "b" (donde Z = a+bY), como hemos indicado ($\hat{a}$ y $\hat{b}$).

5. Estimar los puntos de estratificación óptima para Y, realizando el cambio lineal con $\hat{a}$ y $\hat{b}$, desde los puntos de estratificación óptima ya calculados para Z.

6. Tomar $n_h(y) = n_h(z)$ unidades del estrato h de Y, pudiendo aprovechar o reutilizar las observaciones tomadas en la muestra

piloto, completando hasta el tamaño muestral indicado por la afijación mínima u óptima, ya calculada en 3.

7. Estimar $\mu(y)$ ó $\sigma^2(y)$ por los estimadores usuales en muestreo estratificado, que requieren el conocimiento de los pesos relativos $W_h(y) = W_h(z)$, ya calculados en el punto 3.

De este modo queda completada la inferencia a realizar en la práctica, aprovechando las ventajas en precisión que consigue el muestreo estratificado optimizado, aunque implica un cálculo numérico mayor por precisarse los puntos de estratificación óptima, los tamaños muestrales determinados por la afijación mínima ú óptima, así como los tamaños o pesos relativos de los estratos.

Referencias.

[1] Dalenius, T. (1957): *Sampling in Sweden. Contributions to the Methods and Theories of Sample Survey Practice.* Almq-vist and Wicksells. Estocolmo.

[2] Ruiz, M. (1990): *Una clase de estimadores de la media poblacional robustos e invariantes lineales.* Metron **48**, pp. 55-66.

[3] Ruiz, M. y Ruiz, M.M. (1992): *Equilibrated strategy for population variance estimation.* Test **1**, pp. 79-91.

[4] Särndal, C.E., Swensson, B. y Wretman, J.H. (1991): *Model Assisted Survey Sampling.* Springer-Verlag. New York.

Rev. Academia de Ciencias. Zaragoza. 49 (1994)

Distribución del número de unidades distintas en una muestra aleatoria simple con reemplazamiento de una población finita

M. Ruiz Espejo

Departamento de Estadística e Investigación Operativa
Facultad de Ciencia Económicas y Empresariales
Universidad Complutense
28223 - MADRID

Summary.

We determine the probability mass function of the number of different units selected in a simple random sample with replacement, of size n (≥ 1), of a finite population of size N.

1. Introduction.

Dada una población finita de tamaño N, se toma una muestra aleatoria simple con reemplazamiento de tamaño n(≥ 1). El número ν ($= 1,2,3,...,\min\{n,N\}$) de unidades distintas seleccionadas en la muestra es una variable aleatoria discreta cuya función de cuantía vamos a calcular.

Esta variable aleatoria ha sido citada por Cassel et al. (1977) y, recientemente, por Thompson (1992), dando algunas propiedades de la misma como que la media muestral, basada en las ν unidades distintas seleccionadas en la muestra aleatoria simple con reemplazamiento, es un estimador insesgado de la media poblacional.

Sin embargo no conocemos ningún trabajo en el que figure de forma expresa su función de cuantía. Este es el objetivo de la presente nota

2. Función de cuantía.

Aplicando la regla de Laplace, para $\nu = 1,2,...,m = \min\{n,M\}$, la función de cuantía es la siguiente

$$p(\nu) = \frac{\binom{N}{\nu}}{N^n} \sum_{i_1=1}^{n-\nu+1} \sum_{i_2=i_1+1}^{n-\nu+2} \cdots \sum_{i_{\nu-1}=i_{\nu-2}+1}^{n-1} PR_n^{i_1, i_2-i_1, \ldots, i_{\nu-1}-i_{\nu-2}, n-i_{\nu-1}}$$

donde

$$PR_m^{m_1,m_2,\ldots,m_k} = \frac{m!}{m_1!\, m_2! \ldots m_k!}$$

si

$$m = \sum_{i=1}^{k} m_i$$

es el número de permutaciones con repetición de m elementos tomados de m_1 en m_1 iguales, m_2 en m_2 iguales,...hasta m_k en m_k iguales.

Referencias

(1) Cassel, C.M., Särndal,C.E.y Wretman, J.H.(1977).*Foundations of Inference in Survey Sampling.* New York: Wiley.

(2) Thompson, S.K. (1992). *Sampling.* New York: Wiley.

Rev. Academia de Ciencias. Zaragoza. 49 (1994)

On Stochastic Global Smoothness

George A. Anastassiou

Dept. of Mathematical Sciences, Memphis State University
Memphis, Tennessee 38152, USA

and

Heinz H. Gonska

Dept. of Mathematics, University of Duisburg
D-47048 Duisburg, Germany

Abstract

Let $(\Omega, \mathcal{A}, P)$ be a probability space and let $C_\Omega[a,b]$ denote the space of stochastically continuous stochastic processes with index set $[a,b]$. When $C[a,b] \subset V \subset C_\Omega[a,b]$ and $\widetilde{L}: V \to C_\Omega[a,b]$ is an E(expectation)-commutative linear operator on V, sufficient conditions are given for E-preservation of global smoothness of $X \in V$ through $\widetilde{L}$. Namely, it is proved that

$$\omega_1(E(\widetilde{L}X); \delta) \leq \|L\| \cdot \widetilde{\omega}_1\left(EX; \frac{c \cdot \delta}{\|L\|}\right) \leq (\|L\| + c) \cdot \omega_1(EX; \delta), \text{ where}$$

$L := \widetilde{L}|_{C[a,b]}$, and for $0 \leq \delta \leq b - a$, ω_1 denotes the first order modulus of continuity with $\widetilde{\omega}_1$ its least concave majorant and c a universal constant. Applications are given to different types of stochastic convolution operators defined through a kernel. In particular, are studied extensively in this connection, stochastic operators defined through a bell-shaped trigonometric kernel. Another application of the above result is to stochastic discretely defined Kratz and Stadtmüller operators.

1991 Mathematics Subject Classification: 41A17, 26A15, 26A18, 60G99.

Key Words: global smoothness preservation, modulus of continuity, stochastic processes, stochastic approximation, convolution-type operators, bell-shaped kernels, stochastic Bernstein operators, discretely defined operators.

1. Introduction

In approximating a stochastic process $X = X(t,\omega)$ by means of approximation operators $\tilde{L}_n$, it is interesting to find out which properties of X are preserved by the approximants $\tilde{L}_n X$. For instance, one can be interested in comparing global smoothness characteristics of X and $\tilde{L}_n X$. Global smoothness of a stochastically continuous stochastic process $X(t,\omega)$ can be expressed by the behaviour of the modulus of continuity $\omega_1(EX;\cdot)$, where E is the expectation operator.

A study of the convergence of monotone linear operators $\tilde{L}_n$, defined on a space of stochastic processes, to the unit operator has been done more recently by M. Weba [16, 17] and G.A. Anastassiou [2].

In the present note we give the stochastic analogue of our main theorem in [3] (see also [4]) and apply it to various types of operators such as the stochastic analogues of transformed convolution-type operators investigated recently by Cao and Gonska [7,8], of a class of discretely defined operators as considered earlier by Kratz and Stadtmüller [11], and of a further class of convolution-type operators dealt with by, among others, Anastassiou [1].

The results of this note show that the stochastic analogues of many approximation operators quite naturally inherit qualitative properties of their non-stochastic predecessors. Furthermore, Section 4.1 below (dealing with global smoothness preservation by transformed convolution-type operators) seems to be of interest in itself in the sense that global smoothnes is preserved by slight modifications of the most powerful approximation operators.

2. Preliminaries

Let $(\Omega,\mathcal{A},P)$ denote a fixed probability space and $L^1(\Omega,\mathcal{A},P)$ the set of all $(\Omega,\mathcal{A})$-$(\mathbb{R},\mathcal{B})$ measurable mappings $Z = Z(\omega)$ with $\int_\Omega |Z(\omega)| \cdot P(d\omega) < \infty$, where $\mathcal{B}$ is the σ-field of Borel subsets of $\mathbb{R}$. By $X = X(t,\omega)$ we will denote a stochastic process with index set $[a,b]$ and real state space $(\mathbb{R},\mathcal{B})$.

The space of stochastically bounded processes is given by

$$B_\Omega[a,b] := \left\{ X : \sup_{t \in [a,b]} \int_\Omega |X(t,\omega)| \cdot P(d\omega) < \infty \right\} .$$

Note that $B[a,b]$, the space of real-valued and bounded functions on the compact interval $[a,b]$, is a subspace of $B_\Omega[a,b]$. Furthermore, the vector space of stochastic processes being stochastically continuous in the L^1-sense is defined by

$$C_\Omega[a,b] := C([a,b], L^1(\Omega,\mathcal{A},P)) := \{X : \int_\Omega |X(t_n,\omega) - X(t,\omega)| \, P(d\omega) \to 0 \text{ for } t_n \to t\} .$$

A subspace of $C_\Omega[a,b]$ is the space $C^0_\Omega[a,b]$ of all *sample continuous* processes. Here a stochastic process $X(t,\omega)$ defined for t in a topological space is called sample continuous iff, for all $\omega \in \Omega$, the *paths* (partial functions, trajectories) $X(\cdot,\omega) : t \mapsto X(t,\omega)$ are continuous (cf. [9, p. 351]). We thus have the natural inclusions

$$C[a,b] \subset C^0_\Omega[a,b] \subset C_\Omega[a,b] \subset B_\Omega[a,b].$$

3. A Theorem on Stochastic Global Smoothness Preservation

In order to formulate the basic theorem of this note, we need the following auxiliary results.

Lemma 3.1. (see Weba [17, Lemma 2.1 (ii)])

If $X \in C_\Omega[a,b]$, then $EX \in C[a,b]$, where E denotes the expectation operator given by

$$(EX)(t) = \int_\Omega X(t,\omega) \cdot P(d\omega).$$

A linear operator $L: V \to C_\Omega[a,b]$ is said to be *E-commutative* on the subspace V, $C[a,b] \subset V \subset C_\Omega[a,b]$, if $E(LX) = L(EX)$ for all $X \in V$. E-commutative operators leave the space $C[a,b]$ invariant as can be seen from the following

Lemma 3.2.

Let L be E-commutative on V, where $C[a,b] \subset V$. Then L maps $C[a,b]$ into $C[a,b]$.

Proof. Let $f \in C[a,b]$; then $Lf = L(Ef) = E(Lf)$.

Here $Lf \equiv Y(t,\omega)$ for some $Y(t,\omega) \in C_\Omega[a,b]$; i.e., for $t_n \to t$ we have

$$0 \leftarrow \int_\Omega |Y(t_n,\omega) - Y(t,\omega)| \cdot P(d\omega) \geq \left| \int_\Omega (Y(t_n,\omega) - Y(t,\omega)) \cdot P(d\omega) \right|$$

$$= |(EY)(t_n) - (EY)(t)|.$$

Thus $(EY)(t)$ is continuous in t, which means that $E(Lf)$ is continuous in t, and this implies that $Lf = E(Lf) \in C[a,b]$. □

Remark 3.3.

Let $X \in C_\Omega[a,b]$ and $[x_1,x_2] \subset [a,b]$. We have $|X| \in C_\Omega[a,b]$, and thus $E|X| \in C[a,b]$ by Lemma 2.1 of [17]. Hence $\int_a^b (E|X|)(s)\, ds < \infty$, so that Fubini's theorem implies

$$E\left(\int_{x_1}^{x_2} X(s,\omega)\,ds\right) = \int_{x_1}^{x_2} (EX)(s)\,ds\,. \qquad \square$$

Theorem 3.4.

Let V be a subspace of $C_\Omega[a,b]$ such that $C[a,b] \subset V \subset C_\Omega[a,b]$. Let $\widetilde{L} : V \to C_\Omega[a,b]$ be linear and such that the following hold:

(i) $\widetilde{L}$ is E-commutative on V,

(ii) The restriction $L := \widetilde{L}|_{C[a,b]}$ (mapping $C[a,b]$ into itself) is bounded with norm $\|L\| \neq 0$,

(iii) $L : C^1[a,b] \to C^1[a,b]$ such that $\|(Lg)'\|_\infty \leq c \cdot \|g'\|_\infty$ for all $g \in C^1[a,b]$.

Then for all $X \in V$ and $0 \leq \delta \leq b-a$ one has

$$\omega_1(E(\widetilde{L}X);\delta) \leq \|L\| \cdot \tilde{\omega}_1\left(EX, \frac{c\delta}{\|L\|}\right) \leq (\|L\| + c) \cdot \omega_1(EX;\delta)\,.$$

Here $\tilde{\omega}_1$ denotes the least concave majorant of ω_1 with respect to the second variable.

Proof. Apply Corollary 6 of [3] for $f = EX$, also noting that

$$\omega_1(E(\widetilde{L}X);\delta) = \omega_1(L(EX);\delta)\,. \qquad \square$$

Remark 3.5.

(i) The inequalities in Theorem 3.4 remain valid if $\|L\|$ is replaced by any upper bound d.

(ii) Theorem 3.4 is also true for operators $\widetilde{L} : V \to C_\Omega[c,d]$ where $[c,d] \subset [a,b]$. Of course, for this case the assumptions (ii) and (iii) have to be modified appropriately, and a suitable generalization of Lemma 3.2 has to be used.

4. Applications

4.1. Stochastic Convolution-type Operators on $C^0_\Omega[a,b]$

In this section we investigate the stochastic analogues of certain convolution-type operators which play an important role in the approximation of continuous functions by algebraic polynomials (see, e.g., the recent papers [7, 8]). The results presented here are not only of interest in the context of stochastic approximation, but are also new for the 'classical' case.

Suppose that $L : C[a,b] \to C[a,b]$ is a linear operator. For $X \in C^0_\Omega[a,b]$ define

$$(\widetilde{L}X)(t,\omega) := L[X(\cdot,\omega);t].$$

Due to the fact that the stochastic process X is sample continuous, i.e., each

path $X(\cdot,\omega)$ is continuous in t, the above right hand side is well defined for each fixed $\omega \in \Omega$. In this section, w.l.o.g., we shall consider the case $[a,b] = [-1,1]$. Furthermore, we assume that the operator L is given by

$$L(f;x) := \pi^{-1} \int_{-\pi}^{\pi} f(\cos s) \cdot K(s - \arccos x)\, ds ,$$

where the kernel $K \neq 0$ is continuous and 2π-periodic.

We show next that, under some additional assumptions to be made below, the corresponding operator $\tilde{L}$ satisfies the conditions of Theorem 3.4.

Firstly, $\tilde{L}$ indeed maps $C_\Omega^0[a,b]$ into $C_\Omega[a,b]$. To see this, let (t_n) be a sequence in $[-1,1]$ such that t_n converges to t. Then

$$\pi \cdot \int_\Omega |\tilde{L}(X(\cdot,\omega);t_n) - \tilde{L}(X(\cdot,\omega);t)|\, P(d\omega)$$

$$= \int_\Omega \Big| \int_{-\pi}^{\pi} X(\cos s,\omega)\, K(s - \arccos t_n)\, ds$$

$$- \int_{-\pi}^{\pi} X(\cos s,\omega)\, K(s - \arccos t)\, ds \Big|\, P(d\omega)$$

$$\leq \int_\Omega \int_{-\pi}^{\pi} |X(\cos s,\omega)| \cdot |K(s - \arccos t_n) - K(s - \arccos t)|\, ds\, P(d\omega)$$

$$\leq \int_\Omega \int_{-\pi}^{\pi} |X(\cos s,\omega)| \cdot \varepsilon\, ds\, P(d\omega) \quad \text{for } n \geq N(\varepsilon) \text{ by the uniform continuity of } K$$

$$= \varepsilon \cdot \int_{-\pi}^{\pi} \Big\{ \int_\Omega |X(\cos s,\omega)|\, P(d\omega) \Big\}\, ds \quad \text{by Fubini's theorem}$$

$$= \varepsilon \cdot \int_{-\pi}^{\pi} E(|X(\cos s, \cdot)|)\, ds = \varepsilon \cdot c \quad \text{for some } c < \infty.$$

Hence $\pi \cdot \int_\Omega |\tilde{L}(X(\cdot,\omega);t_n) - \tilde{L}(X(\cdot,\omega);t)|\, P(d\omega)$ tends to 0 as t_n approaches t, showing that $\tilde{L}X \in C_\Omega[a,b]$.

We show next that the three additional assumptions of Theorem 3.4 are fulfilled.

(i) $\tilde{L}$ is E-commutative on $V = C_\Omega^0[a,b]$. Indeed,

$$[\tilde{L}(EX)](t,\omega) = \tilde{L}(EX;t,\omega)$$

$$= L[EX(\cdot,\omega);t]$$

$$= \pi^{-1}\int_{-\pi}^{\pi} (EX)(\cos s)\cdot K(s - \arccos t)\, ds$$

$$= \pi^{-1}\int_{-\pi}^{\pi}\int_{\Omega} X(\cos s,\omega)\, P(d\omega)\cdot K(s - \arccos t)\, ds$$

$$= \int_{\Omega}\left\{\pi^{-1}\int_{-\pi}^{\pi} X(\cos s,\omega)\cdot K(s - \arccos t)\, ds\right\} P(d\omega) \quad \text{(Fubini)}$$

$$= E[L(X(\cdot,\omega);t)] = E[\tilde{L}(X;t,\omega)]$$

$$= [E\,\tilde{L}X]\,(t,\omega)\,.$$

Note that, for the above application of Fubini's theorem, we have used the fact that if $X \in C_\Omega[a,b]$, $f \in C[a,b]$, then $f \cdot X \in C_\Omega[a,b]$.

(ii) It is well known that $\|L\| = \pi^{-1}\,\|K\|_{L_1[-\pi,\pi]}$. The assumption that $K \neq 0$ is continuous, implies $\|K\|_{L_1[-\pi,\pi]} > 0$, i.e., $\|L\| \neq 0$.

(iii) To verify this we first give a general estimate for $\left|\frac{d}{dx}\, L(f;x)\right|$. Note that the operator L from above can be written as

$$L(f;x) = \pi^{-1}\int_{-\pi}^{\pi} f(\cos s)\cdot K(s - \arccos x)\, ds$$

$$= \pi^{-1}\int_{-\pi}^{\pi} f(\cos(t + \arccos x))\cdot K(t)\, dt$$

$$= \pi^{-1}\int_{-\pi}^{\pi} f(\cos(\arccos x - t))\cdot K(t)\, dt\,.$$

Writing,

$g := f \circ \cos$, $\theta := \arccos x$, L attains the form

$$L(f;x) = \pi^{-1} \cdot \int_{-\pi}^{\pi} g(\theta - t) \cdot K(t)\, dt =: \overline{L}(g;\theta)$$

Note that $\overline{L}(g;\theta)$ is defined for all $g \in C_{2\pi}$ and $\theta \in \mathbb{R}$. From [6, Prop. 1.1.15] we have

$$\frac{d}{d\theta}\, \overline{L}(g;\theta) = \pi^{-1} \cdot \int_{-\pi}^{\pi} \left\{\frac{d}{dt}\, g(\theta - t)\right\} \cdot K(t)\, dt\,.$$

Here,

$$\frac{d\theta}{dx} = -\frac{1}{\sqrt{1 - x^2}}, \quad \text{so that} \quad \frac{1}{d\theta} = \frac{-\sqrt{1 - x^2}}{dx}\,.$$

Hence,

$$\frac{d}{dx}\, \overline{L}[f \circ \cos, \arccos x] \cdot \sqrt{1 - x^2}$$

$$= -\frac{d}{d\theta}\, \overline{L}[f \circ \cos, \arccos x]$$

$$= -\pi^{-1} \int_{-\pi}^{\pi} \frac{d}{dt}\, f(\cos(\theta - t)) \cdot K(t)\, dt$$

$$= \pi^{-1} \cdot \int_{-\pi}^{\pi} \sin(\theta - t) \cdot f'[\cos(\theta - t)](-1) \cdot K(t)\, dt$$

$$= \pi^{-1} \int_{\theta + \pi}^{\theta - \pi} \sin s \cdot f'[\cos s] \cdot K(\theta - s)\, ds$$

$$= -\pi^{-1} \int_{-\pi}^{\pi} \sin s \cdot f'[\cos s] \cdot K(\theta - s)\, ds$$

$$= -\frac{1}{\pi} \cdot \left(\int_{-\pi}^{0} + \int_{0}^{\pi}\right) \sin s \cdot f'[\cos s] \cdot K(\theta - s)\, ds$$

$$= -\frac{1}{\pi} \int_{0}^{\pi} \sin s \cdot f'(\cos s) \cdot [K(\theta - s) - K(\theta + s)]\, ds\,.$$

Thus,

$$\left|\frac{d}{dx}\, \overline{L}(f \circ \cos, \arccos x)\right| \cdot \sqrt{1 - x^2}$$

$$= \left| \frac{1}{\pi} \cdot \int_0^{\pi} \sin s \cdot f'(\cos s) \cdot [K(\theta - s) - K(\theta + s)]\, ds \right|$$

$$\leq \|f'\| \cdot \frac{1}{\pi} \cdot \int_0^{\pi} \sin s \cdot |K(\theta - s) - K(\theta + s)|\, ds .$$

It thus remains to give a representation of

$$\pi^{-1} \int_0^{\pi} \sin s \cdot |K(\theta - s) - K(\theta + s)|\, ds$$

To this end, first recall that a function $g \in C_{2\pi}$ is called *bell-shaped on* $[-\pi, \pi]$, if it is even and decreases on $[0,\pi]$ (see [12]). Furthermore, it is known from a lemma of Beatson [5] that $g \in C_{2\pi}$ is bell-shaped if and only if for all $t, x \in [0,\pi]$ one has

$$g(t - x) - g(t + x) \geq 0.$$

Thus bell-shaped kernels K constitute an important class of kernels for which the above integral can be further simplified. A whole class of examples will be given below. Indeed for them the above quantity becomes

$$\pi^{-1} \int_0^{\pi} \sin s \cdot [K(\theta - s) - K(\theta + s)]\, ds$$

$$= \frac{1}{2\pi} \cdot \int_{-\pi}^{\pi} \sin s \cdot [K(\theta - s) - K(\theta + s)]\, ds$$

$$= \frac{1}{2\pi} \cdot \left\{ -\int_{+\pi}^{-\pi} \sin(\theta - \tilde{s}) \cdot K(\tilde{s})\, d\tilde{s} - \int_{-\pi}^{\pi} \sin(\tilde{s} - \theta) \cdot K(\tilde{s})\, d\tilde{s} \right\}$$

$$= \frac{1}{2\pi} \cdot \left\{ \int_{-\pi}^{\pi} [\sin(\theta - s) - \sin(s - \theta)] \cdot K(s)\, ds \right\}$$

$$= \frac{1}{\pi} \cdot \int_{-\pi}^{\pi} \sin(\theta - s) \cdot K(s)\, ds$$

$$= \frac{1}{\pi} \cdot \int_{-\pi}^{\pi} [\sin\theta \cdot \cos s - \cos\theta \cdot \sin s] \cdot K(s)\, ds$$

$$= \frac{1}{\pi} \left[\sin\theta \cdot \int_{-\pi}^{\pi} \cos s \cdot K(s)\, ds - \cos\theta \cdot \int_{-\pi}^{\pi} \sin s \cdot K(s)\, ds \right]$$

$$= \frac{1}{\pi} \cdot \sin\theta \cdot \int_{-\pi}^{\pi} \cos s \cdot K(s)\, ds$$

$= : \sin\theta \cdot \varrho_1[K]$.

Thus we have

$$\left|\frac{d}{dx}\, L(f\,;\,x)\right| \sqrt{1-x^2}$$

$$= \left|\frac{d}{dx}\, \bar{L}(f\circ\cos\,,\,\arccos x)\right| \cdot \sqrt{1-x^2} \;\le\; \|f'\| \cdot \sin\theta \cdot \varrho_1[K]\,,$$

or

$$\left|\frac{d}{dx}\, L(f\,;\,x)\right| \;\le\; \varrho_1[K] \cdot \|f'\|\,.$$

Recalling further that for operators L of the form before one has $\|L\| = \pi^{-1} \cdot \|K\|_{L_1[-\pi,\pi]}$, the above can be summarized as follows.

Theorem 4.1.

Let L be a convolution-type operator of the form given above which is based upon the non-negative and bell-shaped kernel $K \neq 0$. Then for all $X \in C^0_\Omega[-1,1]$ and all $0 \le \delta \le 2$ one has

$$\omega_1\big(E(\tilde{L}X)\,;\,\delta\big) \;\le\; \frac{1}{\pi}\|K\|_{L_1[-\pi,\pi]}\, \tilde{\omega}_1\left(EX\,;\,\frac{\varrho_1[K]\cdot\delta}{\pi^{-1}\cdot\|K\|_{L_1[-\pi,\pi]}}\right)$$

$$\le \big[\pi^{-1}\cdot\|K\|_{L_1[-\pi,\pi]} + \varrho_1[K]\big]\cdot\omega_1(EX\,;\,\delta)\,.$$

We now specialize K further by assuming that

$$K(t) = K_{m(n)}(t) = \frac{1}{2} + \sum_{k=1}^{m(n)} \varrho_{k,m(n)} \cdot \cos kt$$

is a non-negative, even and bell-shaped trigonometric polynomial of degree $\le m(n)$. The operators L based upon these kernels will be denoted by $G_{m(n)}$. Then we have

Theorem 4.2.

Let $G_{m(n)}$ be a convolution-type operator as given above. Then for all $X \in C^0_\Omega[-1,1]$ and all $0 \le \delta \le 2$ one has

$$\omega_1\big[E(\tilde{G}_{m(n)}X)\,;\,\delta\big] \;\le\; \tilde{\omega}_1(EX\,;\,\varrho_{1,m(n)}\cdot\delta) \;\le\; \big[1+\varrho_{1,m(n)}\big]\cdot\omega_1(EX\,;\,\delta)\,.$$

Proof. We note first that

$\|K_{m(n)}\|_{L_1[-\pi,\pi]} = \pi$ (so that $\|G_{m(n)}\| = 1$).

Furthermore,

$$\vartheta_1[K_{m(n)}] = \frac{1}{\pi} \cdot \int_{-\pi}^{\pi} \cos s \cdot K_{m(n)}(s)\, ds$$

$$= \frac{1}{\pi} \cdot \int_{-\pi}^{\pi} \cos s \cdot \left(\frac{1}{2} + \sum_{k=1}^{m(n)} \vartheta_{k,m(n)} \cos ks \right) ds$$

$$= \frac{1}{\pi} \cdot \int_{-\pi}^{\pi} \cos^2 s \cdot \vartheta_{1,m(n)}\, ds$$

$$= \frac{1}{\pi} \cdot \vartheta_{1,m(n)} \cdot \left[\frac{1}{2} s + \frac{1}{4} \cdot \sin 2s \right]_{-\pi}^{\pi}$$

$$= \vartheta_{1,m(n)} .$$

The inequalities of Theorem 4.2. then follow directly from Theorem 4.1. □

Corollary 4.3.

Under the above assumptions on $K_{m(n)}$ it can be easily verified that $0 \le \vartheta_{1,m(n)} \le 1$. Thus we have

$$\omega_1\left(E(\tilde{G}_{m(n)}X)\,;\,\delta\right) \le \tilde{\omega}_1(EX\,;\,\delta) \le 2 \cdot \omega_1(EX\,;\,\delta).$$

Example 4.4. (Construction of bell-shaped kernels)

In Beatson's report [5] the author constructs positive and bell-shaped kernels using Steklov means. To be more specific, let $r \in \mathbb{N} = \{1, 2, \dots\}$ and

$$J_{rn-r}(t) = c_n \cdot \left[\frac{\sin \frac{nt}{2}}{\sin \frac{t}{2}} \right]^{2r} = \frac{1}{2} + \sum_{k=1}^{rn-r} \vartheta_{1,rn-r} \cdot \cos kt .$$

Here the constant c_n is chosen so that $\frac{1}{\pi} \cdot \int_{-\pi}^{\pi} J_{r(n-1)}(t)\, dt = 1$.

For $r = 1$ we obtain the Fejér kernel, $r = 2$ gives the Jackson kernel, and for

$r \geq 3$ one arrives at Jackson kernels of higher order (Matsuoka kernels). Beatson defined new kernels $\varnothing_{rn-r}(x)$, based upon the J_{rn-r}, and given as

$$\varnothing_{rn-r}(x) := \frac{n}{2\pi} \cdot \int_{-\pi/n}^{\pi/n} J_{rn-r}(x+t)\,dt =: \frac{1}{2} + \sum_{k=1}^{rn-r} \lambda_{k,rn-r} \cdot \cos kt \ .$$

He noted that these are bell-shaped. It remains to be shown what $\lambda_{1,rn-r}$ looks like for these kernels. For the higher order Jackson kernels J_{rn-r} as given above one has

$$\varrho_{1,rn-r} = \frac{1}{\pi} \int_{-\pi}^{\pi} \cos s \cdot J_{rn-r}(s)\,ds \ .$$

Below we show the relationship between $\varrho_{1,rn-r}$ and $\lambda_{1,rn-r}$. Note that for the argument we only need the fact that J_{rn-r} is even. One has

$$\lambda_{1,rn-r} = \frac{1}{\pi} \cdot \int_{-\pi}^{\pi} \cos s \cdot \left\{ \frac{n}{2\pi} \int_{-\pi/n}^{\pi/n} J_{rn-r}(s+t)\,dt \right\} ds$$

$$= \frac{n}{2\pi^2} \int_{-\pi/n}^{\pi/n} \left\{ \int_{-\pi}^{\pi} \cos s \cdot J_{rn-r}(s+t)\,ds \right\} dt \quad \text{(Fubini)}$$

$$= \frac{n}{2\pi^2} \int_{-\pi/n}^{\pi/n} \left\{ \int_{-\pi}^{\pi} \cos(s-t) \cdot J_{rn-r}(s)\,ds \right\} dt$$

Using the trigonometric identity $\cos(r-t) = \cos r \cdot \cos t + \sin r \cdot \sin t$ it is seen that

$$\lambda_{1,rn-r} = \frac{n}{2\pi^2} \cdot \int_{-\pi/n}^{\pi/n} \left\{ \int_{-\pi}^{\pi} (\cos s \cdot \cos t + \sin s \cdot \sin t) \cdot J_{rn-r}(s)\,ds \right\} dt$$

$$= \frac{n}{2\pi^2} \cdot \int_{-\pi/n}^{\pi/n} \left\{ \cos t \cdot \int_{-\pi}^{\pi} \cos s \cdot J_{rn-r}(s)\,ds \right.$$

$$\left. + \sin t \cdot \int_{-\pi}^{\pi} \sin s \cdot J_{rn-r}(s)\,ds \right\} dt$$

where the second summand equals 0 due to J_{rn-r} being even. Hence,

$$\lambda_{1,rn-r} = \frac{n}{2\pi^2} \cdot \int_{-\pi/n}^{\pi/n} \cos t \cdot \pi \cdot \varrho_{1,rn-r}\, dt$$

$$= \varrho_{1,rn-r} \cdot \frac{n}{2\pi} \cdot \int_{-\pi/n}^{\pi/n} \cos t\, dt$$

$$= \varrho_{1,rn-r} \cdot \frac{n}{\pi} \cdot \sin \frac{\pi}{n} .$$

Thus our conclusion for the operators L based upon Beatson's kernels φ_{rn-r}, which we now denote by W_{rn-r}, can be summarized as follows.

Theorem 4.5.

Let W_{rn-r} be the convolution-type operator based upon φ_{rn-r}, where φ_{rn-r} denotes Beatson's modification of the Jackson kernel J_{rn-r}, $r \geq 1$. Then for all $X \in C_{\Omega}^{0}[-1,1]$ and all $0 \leq \delta \leq 2$ there holds

$$\omega_1(E(\widetilde{W_{rn-r}}X)\,;\,\delta) \leq \tilde{\omega}_1(EX\,;\,\lambda_{1,rn-r} \cdot \delta) \leq \tilde{\omega}_1(EX\,;\,\delta)\,,$$

and also

$$\omega_1(E(\widetilde{W_{rn-r}}X)\,;\,\delta) \leq \tilde{\omega}_1(EX\,;\,\lambda_{1,rn-r} \cdot \delta) \leq (1 + \lambda_{1,rn-r}) \cdot \omega_1(EX\,;\,\delta) \leq 2 \cdot \omega_1(EX\,;\,\delta)\,.$$

To conclude this section, below we give explicit representations of $\varrho_{1,rn-r}$ for $r = 1, 2, 3, 4$. These can be found in the literature, but sometimes in less accessible sources. General, however less explicit representations for all convergence factors $\varrho_{k,rn-r}$ of higher order Jackson kernels can be found in a paper by Matsuoka [13].

Example 4.6.

Here we give explicit representations of the convergence factors $\varrho_{1,rn-r}$ of the Jackson kernels J_{rn-r} for $r = 1, 2, 3, 4$.

Fejér kernel ($r = 1$):

In this case one has $\varrho_{1,n-1} = 1 - \frac{1}{n}$ [see, e.g., [15, p. 69]].

Jackson kernels of higher orders ($r \geq 2$):

The convergence factors $\varrho_{k,r(n-1)}$ of J_{rn-r} can be written as [see [10, p. 37 f.]]

$$\varrho_{k,r(n-1)} = \frac{\mu_{k,n}(r,r)}{\mu_{0,n}(r,r)}\,, \quad \text{where}$$

$$\mu_{k,n}(r,r) = \sum_{j=0}^{2r} (-1)^j \binom{2r}{j} \binom{nr + r - 1 - jn - k}{2r - 1}, \quad 0 \le k \le r(n-1).$$

Note that in the latter representation the convention $\binom{n}{k} := 0$ for $n,k \in \mathbf{Z}$, $n < k$, is used.

A simple computation for the case $r = 2$ shows that

$$\mu_{1,n}(2,2) = \frac{2}{3} n(n^2 - 1) \text{ and}$$

$$\mu_{0,n}(2,2) = \frac{1}{3} n(2n^2 + 1), \text{ so that}$$

$$\vartheta_{1,2(n-1)} = 1 - \frac{3}{2n^2 + 1} \quad \text{[cf. [6, p. 60]].}$$

For $r = 3$ it was shown by Stark [14, p. 73] that

$$\vartheta_{1,3(n-1)} = \frac{11n^4 - 5n^2 - 6}{11n^4 + 5n^2 + 4},$$

and for $r = 4$ one has the explicit representation [[14, p. 74]]

$$\vartheta_{1,4(n-1)} = 1 - \frac{105(n^4 + n^2 + 1)}{151n^6 + 70n^4 + 49n^2 + 45}.$$

From these the corresponding factors $\lambda_{1,rn-r}$ of φ_{rn-r} can be easily derived.□

4.2. Operators on $C_\Omega[a,b]$

If one wants to investigate global smoothness preservation of more general stochastic processes, i.e., $X \in C_\Omega[a,b]$, it is appropriate to consider special approximation operators L mapping $C[a,b]$ into itself. In [3] we investigated global smoothness preservation by discretely defined operators of the form [see Kratz and Stadtmüller [11]]

$$L_n(f;x) = \sum_{j \in J_n} f(\xi_{j,n}) \cdot p_{j,n}(x), \quad \text{where } J_n \text{ is a finite index set.}$$

Furthermore, we assumed that the following hold:

(a) $\sum_{j \in J_n} p_{j,n}(x) \equiv s_n \neq 0$,

(b) $\sum_{j \in J_n} |p_{j,n}(x)| \le c_1 \neq 0$,

(c) $p_{j,n} \in C^1[a,b]$ and $\sum_{j \in J_n} |(\xi_{j,n} - x) \cdot p'_{j,n}(x)| \le c_2$.

For these we have by Theorem 10 in [3] that

$$\omega(L_n f;t) \le c_1 \cdot \tilde{\omega}_1\left(f ; \frac{c_2 t}{c_1}\right) \le (c_1 + c_2) \cdot \omega_1(f;t).$$

The classical Bernstein operators are typical examples of operators with these properties.

Recall the representation

$$\tilde{L}_n X(t,\omega) = L_n[X(\cdot,\omega);t]$$

$$= \sum_{j \in J_n} X(\xi_{j,n},\omega) \cdot p_{j,n}(t)$$

We show next that Theorem 3.4 can be applied with $V = C_\Omega[a,b]$. First we verify that $\tilde{L}_n : C_\Omega[a,b] \to C_\Omega[a,b]$. Let $t_N \to t$. Then

$$\int_\Omega \left|\tilde{L}_n X(t_N,\omega) - \tilde{L}_n X(t,\omega)\right| P(d\omega)$$

$$= \int_\Omega \left| \sum_{j \in J_n} X(\xi_{j,n},\omega) \cdot \left(p_{j,n}(t_N) - p_{j,n}(t)\right)\right| P(d\omega)$$

$$\le \int_\Omega \sum_{j \in J_n} \left|X(\xi_{j,n},\omega)\right| \cdot \left|p_{j,n}(t_N) - p_{j,n}(t)\right| P(d\omega)$$

$$\le \sum_{j \in J_n} \left|p_{j,n}(t_N) - p_{j,n}(t)\right| \cdot \int_\Omega \left|X(\xi_{j,n},\omega)\right| P(d\omega)$$

$$\le \sum_{j \in J_n} M \cdot \left|p_{j,n}(t_N) - p_{j,n}(t)\right|$$

where $M := \sup_{t \in [a,b]} \int_\Omega |X(t,\omega)| P(d\omega) < +\infty$, since $X \in B_\Omega[a,b]$.

Here, due to the continuity of $p_{j,n}$, $j \in J_n$, all differences $p_{j,n}(t_N) - p_{j,n}(t)$ tend to 0, so that $\tilde{L}_n X \in C_\Omega[a,b]$.

(i) Show that $\tilde{L}_n$ is E-commutative.

$$E\tilde{L}_n X(t,\omega) = E\left(\sum_{j \in J_n} X(\xi_{j,n},\omega) \cdot p_{j,n}(t)\right)$$

$$= \sum_{j \in J_n} EX(\xi_{j,n}) \cdot p_{j,n}(t)$$

$$= [\tilde{L}_n(EX)](t,\omega)$$

(ii) $\|L_n\| \neq 0$, since $s_n \neq 0$. Also, by condition (b) it follows that $\|L_n\| \leq c_1 < \infty$.

(iii) Obviously, $L_n : C^1[a,b] \to C^1[a,b]$. From [3] it follows that

$$\|(Lg)'\| \leq c_2 \cdot \|g'\| \quad \text{for all } g \in C^1[a,b] .$$

Hence we conclude from Theorem 3.4 and Remark 3.5 that

$$\omega_1(E(\tilde{L}_n X);\delta) \leq c_1 \cdot \tilde{\omega}_1\left(EX;\frac{c_2\delta}{c_1}\right) \leq (c_1 + c_2) \cdot \omega_1(EX;\delta)$$

for all $X \in C_\Omega[a,b]$ and $0 \leq \delta \leq b - a$.

As a typical example one may consider stochastic Bernstein operators on $C_\Omega[0,1]$ defined by (see, e.g., [16])

$$[B_n X](t,\omega) = \sum_{k=0}^{n} X\left(\frac{k}{n}, \omega\right) \cdot \binom{n}{k} t^k (1-t)^{n-k} .$$

For these one obtains for all $X \in C_\Omega[0,1]$ and $0 \leq \delta \leq 1$ the inequality

$$\omega_1[E(\tilde{B}_n X);\delta] \leq \tilde{\omega}(EX;\delta) \leq 2 \cdot \omega_1(EX;\delta) .$$

4.3. Further Convolution-type Operators

The following type of stochastic convolution operators obeys the global smoothness preservation property in a natural direct way by producing a better constant in the associated inequality which can be in terms of both the first order modulus of continuity ω_1 and the second order one ω_2, thus improving the inequality coming from our main theorem 3.4. We will discuss both approaches below.

Example 4.7.

Consider the positive linear convolution operators [see [1]]

$$L_n : C[-2a,2a] \to C[-a,a] , \qquad a > 0 ,$$

given by

$$L_n(f,x) = \int_{-a}^{a} \left(\frac{f(x+y) + f(x-y)}{2}\right) g_n(y)\, dy , \quad f \in C[-2a,2a] , \quad n \in \mathbb{N} ,$$

where $g_n(y) \geq 0$ is continuous and $\int_{-a}^{a} g_n(y)\, dy = 1$.

The corresponding stochastic positive linear convolution operators now have the form

$$[\tilde{L}_n(X)](t,\omega) = \int_{-a}^{a} \left(\frac{X(t+y,\omega) + X(t-y,\omega)}{2} \right) \cdot g_n(y)\, dy ,$$

map $C_\Omega[-2a,2a]$ into $C_\Omega[-a,a]$ and are $\neq 0$. Here $t,y \in [-a,a]$. To show that Theorem 3.4 is also applicable in this case, we note the following:

(i) $\tilde{L}_n$ is E-commutative, that is

$$(E\tilde{L}_n X)(t,\omega) = \int_{-a}^{a} \frac{(EX)(t+y) + (EX)(t-y)}{2} \cdot g_n(y)\, dy$$

$$= (\tilde{L}_n(EX))(t,\omega), \text{ i.e., } E\tilde{L}_n = \tilde{L}_n E .$$

(ii) $\tilde{L}_n|_{C[-2a,2a]}$ is a bounded operator, i.e.,

$$|L_n(f,x)| \le \int_{-a}^{a} \frac{|f(x+y) + f(x-y)|}{2}\, g_n(y)\, dy \le 1 \cdot \|f\|_\infty ,$$

in fact $\|L_n\| = 1$.

(iii) Observe that

$L_n: C^1[-2a,2a] \to C^1[-a,a]$ and $\|(Lf)'\|_\infty \le \|f'\|_\infty$, i.e., $c = 1$ in Theorem 3.4 (iii).

Indeed,

$$[L_n(f,x)]' = \int_{-a}^{a} \frac{f'(x+y) + f'(x-y)}{2} \cdot g_n(y)\, dy = L_n(f',x) ,$$

and thus

$$|(L_n(f,x))'| \le \|f'\|_\infty , \text{ that is } \|(L_n f)'\|_\infty \le \|f'\|_\infty .$$

Therefore, from Theorem 3.4 [see also Remark 3.5 (ii)] we immediately get

$$\omega_1[E(\tilde{L}_n X),\delta] \le \tilde{\omega}(EX,\delta) \le 2\omega_1(EX,\delta) , \text{ for all } X \in C_\Omega[-2a,2a] , \ 0 \le \delta \le 2a . \quad \square$$

Remark 4.8.

Let $f \in C[-2a,2a]$. One easily observes that

$$\|L_n f - f\|_\infty \le \frac{1}{2}\int_{-a}^{a} \omega_2(f,|y|)\, g_n(y)\, dy ,$$

that is, for proper choices of the g_n we obtain approximation operators. Also one can establish directly that

$$\omega_1(L_n f,h) \le \omega_1(f,h) \quad \text{and} \quad \omega_2(L_n f,h) \le \omega_2(f,h) ,$$

that is obtain the constant 1 on the right hand side, instead of the 2 which could be derived using the main result of [3]. For $X \in C_\Omega[-2a,2a]$ we have that $EX \in C[-2a,2a]$. Also, from $\tilde{L}_n X \in C_\Omega[-a,a]$ we have $E(L_n X) \in C[-a,a]$. Note also that $E\tilde{L}_n = \tilde{L}_n E$, i.e.,

$$\omega_{1,2}[E(\tilde{L}_n X),\delta] = \omega_{1,2}[\tilde{L}_n(EX),\delta] .$$

Therefore picking $f := EX$ and applying the above inequalities, we obtain

$$\omega_{1,2}[E(\tilde{L}_n X),\delta] \le \omega_{1,2}(EX,\delta) .$$

This is better than what was derived from Theorem 3.4. in the previous example. □

Acknowledgement

Both authors express their sincere thanks to Ms. E. Müller-Faust from EBS for her technical assistance while writing this note.

References

[1] G.A. Anastassiou: Sharp inequalities for convolution-type operators, J. Approx. Theory 58 (1989), 259-266.

[2] G.A. Anastassiou: Korovkin inequalities for stochastic processes. "J. Math. Anal. Appl." 157 (1991), 366-384.

[3] G.A. Anastassiou, C. Cottin, H.H. Gonska: Global smoothness of approximating functions. Analysis 11 (1991), 43-57.

[4] G.A. Anastassiou, C. Cottin, H.H. Gonska: Global smoothness preservation by multivariate approximation operators. In: Israel Mathematical Conference Proceedings, Vol. IV (ed. by S. Baron und D. Leviatan), 31-44. Ramat-Gan: Bar-Ilan University 1991.

[5] R.K. Beatson: Bell-shape preserving convolution operators. Technical Report (1978), Dept. of Math., University of Otago, Dunedin, New Zealand.

[6] P.L. Butzer, R.J. Nessel: Fourier Analysis and Approximation (Vol. I), Birkhäuser, Basel and Stuttgart, 1971.

[7] Jia-ding Cao, H. Gonska: Approximation by Boolean sums of linear operators II: Gopengauz-type estimates. J. Approx. Theory 57 (1989), 77-89.

[8] Jia-ding Cao, H. Gonska: Approximation by Boolean sums of linear operators III: Estimates for some numerical approximation schemes. Numer. Funct. Anal. Optim. 10 (1989), 643-672.

[9] R.M. Dudley: Real Analysis and Probability, Wadsworth & Brooks/Cole, Pacific Grove/CA, 1989.

[10] E. Görlich, E.L. Stark: Über beste Konstanten und asymptotische Entwicklungen positiver Faltungsintegrale und deren Zusammenhang mit dem Saturationsproblem. Jahresber. Deutsch. Math.-Verein. 27 (1970), 18-61.

[11] W. Kratz, U. Stadtmüller: On the uniform modulus of continuity of certain discrete approximation operators. J. Approx. Theory 54 (1988), 326-337.

[12] G.G. Lorentz, K. Zeller: Degree of approximation by monotone polynomials. J. Approx. Theory 1 (1968), 501-504.

[13] Y. Matsuoka: On the approximation of functions by some singular integrals, Tôhoku Math. J. 18 (1) (1966), 13-43.

[14] E.L. Stark: Über trigonometrische singuläre Faltungsintegrale mit Kernen endlicher Oszillation. Dissertation, Technische Hochschule Aachen, 1970.

[15] E.L. Stark: Nikol'skiĭ-Konstanten und Approximationsmaße im Hilbert-Raum. Habilitationsschrift, Technische Hochschule Aachen, 1978.

[16] M. Weba: Korovkin systems of stochastic processes, Math. Z. 192 (1986), 73-80.

[17] M. Weba: A quantitative Korovkin theorem for random functions with multivariate domains. J. Approx. Theory 61 (1990), 74-87.

The research of both authors was supported under *NATO grant CRG. 891013.* They also gratefully acknowledge the hospitality of the Mathematical Research Institute at Oberwolfach where this note was finished during a pleasant one-week stay.

Rev. Academia de Ciencias. Zaragoza. 49 (1994)

UNBIASED ESTIMATORS USING AUXILIARY INFORMATION IN SAMPLE SURVEYS: A REVIEW

L. N. Sahoo and M. Ruiz Espejo

Department of Statistics
Utkal University
Bhubaneswar 751004
INDIA

and

Departamento de Estadística
Facultad de Ciencias Económicas y Empresariales
Universidad Complutense
28223 Madrid
ESPAÑA

Summary

One of the methods of increasing the precision of the estimates in sample surveys is the use of auxiliary information. But, the main drawback of the commonly adopted estimators exploiting auxiliary information on one (or more than one) auxiliary variable is that they are biased. This as prompted many research workers to develop estimators that are either almost unbiased (weekly biased) or wholly unbiased. This paper presents a brief review of research work on some unbiased estimators recently developed in sample survey theory for sampling from finite populations.

1. Introduction and some preliminaries.

Let y and x denote the study variable and the auxiliary variable taking values y_i and x_i $(i = 1,2,\ldots,N)$ respectively on the ith unit of a finite population. A simple random sample (WOR) of size n $(n < N)$ is selected in orden to estimate the population mean $\overline{Y}$ of y, when the population mean $\overline{X}$ of x is known.

It is well known that when x has a high positive correlation with y, one can use the ratio method of estimation and obtain two reasonable "design biased" ratio estimators

$$\overline{y}_R = r\overline{X} \quad \text{and} \quad \overline{y}'_R = \overline{r}\overline{X} \tag{1}$$

where $r = \frac{\overline{y}}{\overline{x}}$; $\overline{y}$, $\overline{x}$ and $\overline{r}$ are the simple arithmetic means of the sample of y_i, x_i and $r_i = \frac{y_i}{x_i}$ respectively. If x has a high negative correlation with y, a natural analogue of the ratio method of estimation is the product method of estimation, which results in the design-biased product estimators

$$\overline{y}_P = \frac{p}{\overline{X}} \quad \text{and} \quad \overline{y}'_P = \frac{\overline{p}}{\overline{X}}, \tag{2}$$

where $p = \overline{y}\,\overline{x}$ and $\overline{p}$ is the simple arithmetic mean of the sample of $p_i = y_i x_i$. However, the design biased linear regression estimator

$$\overline{y}_{RG} = \overline{y} - b_{yx}\left(\overline{x} - \overline{X}\right), \tag{3}$$

where b_{yx} is the sample regression coefficient of y on x, can be used for both the situations of positively and negativaly correlated variables. The precision of $\overline{y}_{RG}$ is usually higher than that of $\overline{y}_R$ and $\overline{y}'_R$ or $\overline{y}_P$ and $\overline{y}'_P$. But, in the large scale surveys these estimators are frequently used for their simplicity.

In many surveys, data on p $(p > 1)$ auxiliary x-variates $(x_1, x_2, \ldots, x_p)$ are available. It is then natural to investigate whether data on all the auxiliary variables can be used to provide an efficient estimator of $\overline{Y}$. Olkin (1958), introduced multivariate ratio method of estimation which later on extended to the multivariate product and regresssion methods of estimation by Singh (1967 a) and Srivastava (1965) respectively. When q $(q < p)$ auxiliary variables $(x_1, x_2, \ldots, x_q)$ are positively and $(p - q)$ auxiliary variables $(x_{q+1}, x_{q+2}, \ldots, x_p)$ are negatively correlated with y, Rao an Mudholkar (1967) have extended Olkin's multivariate ratio estimator to a weighted combination of ratio and product estimators. But, in the same situation, Singh (1967 b) has suggested an estimator called ratio-cum-product estimator.

Lack of unbiasedness in the classical estimator has encouraged many research workers to develop unbiased estimators or by modifying the basic designs.

2. Unbiased estimators using single auxiliary variable

Basic work on unbiased ratio estimation was initiated by Hartley and Ross (1954). They constructed an unbiased ratio estimator

$$\overline{y}_{HR} = \overline{y}'_R + \frac{(N-1)n}{N(n-1)}\left(\overline{y} - \overline{r}\,\overline{x}\right) \tag{1}$$

starting with $\overline{y}'_R$ and correcting it for the bias. Goodman an Hartley (1958) derived the exact variance of $\overline{y}_{HR}$ when $N >> n$, and studied the relative precision of $\overline{y}_R$ and $\overline{y}_{HR}$. Mickey (1959) and Williams (1961) have constructed broad classes of unbiased estimators

from which the $\overline{y}_{HR}$ comes out as a particular case. Ruiz and Santos (1989) offered two expressions of the bias of $\overline{y}'_R$ and proposed a new class of unbiased estimators including $\overline{y}_{HR}$ and based on the same or less statistics, as

$$\overline{y}_{RS} = \frac{n(N-1)}{N-n}\overline{r}\,\overline{x} - \frac{(n-1)N}{N-n}\overline{y}'_R, \tag{2}$$

which does not include $\overline{y}$.

In practice, it is not possible to obtain an unbiased estimator of the bias of $\overline{y}_R$ under simple random sampling. So, many authors have studied the problem of constructing almost unbiased ratio estimators (unbiased upto terms of order n^{-1}) by considering suitable modifications of $\overline{y}_R$. The notable ones in this direction are due to Quenouille (1956), Murthy and Nanjamma (1959), Pascual (1961), Beale (1962), Tin (1965), Sahoo (1983, 1987), Singh, Iachan and Upadhyaya (1985) among others. But, however by correcting the sampling design, many authors have managed to make $\overline{y}_R$ a completely unbiased estimator. Lahiri (1951) showed that $\overline{y}_R$ is unbiased if the sample is drawn with probability proportional to $\sum x_i$. The simplest method of doing this is due to Midzuno (1952) and Sen (1952) in which the first member of the sample is draw with probability proportional to x_i and rest $(n-1)$ units with SRSWOR. Midzuno's technique of changing the selection procedure for obtainig unbiased ratio estimators was further studied by Raj (1954), and Nanjamma, Murthy and Sethi (1959). Another generalisation of the Lahiri's method and the corresponding unbiased ratio estimator was given by Deshpande (1984).

During the years that followed several attempts were also made to construct unbiased product estimators which run parallel to the construction of the unbiased ratio estimators. The early contributions in this direction were due to Robson (1957) and Murthy (1964). Srivastava, Shukla and Bhatnagar (1981) have shown that Robson's estimator gives a better performance than Murthy's estimator. Srivenkataramana and Tracy (1979) reviewed some of these methods while at the same time offering some more estimates. Vos (1980) generalized these methods to obtain mixing estimators and considered the efficiency of these estimators. In the recent past, unbiased product estimators were also considered by Shah and Shah (1979), Gupta and Adhvaryu (1982), Iachan, Singh and Upadhyaya (1987), Rao (1983, 1987), Singh (1989).

For the first time Ruiz and Santos (1990) proposed a new sampling design for which a new product estimator is rendered unbiased. Under this design the probabililty of selecting s–th sample is given by

$$p(s) = \frac{m}{\binom{N}{n}\overline{x}} \tag{3}$$

and the estimator taken into consideration is $\overline{y}_{PR} = \frac{\overline{y}\,\overline{x}}{m}$, where m is the equiprobable harmonic mean of $\overline{x}$–values.

More recently, Dalabehera and Sahoo (1993) generated three unbiased estimators by correcting the estimators $\overline{y}\frac{\overline{X}_h}{\overline{x}_h}$, $\overline{r}\overline{X}_h$ and $\frac{\overline{p}}{\overline{X}_h}$ for their biases, where $\overline{x}_h$ and $\overline{X}_h$ are respectively the sample and population harmonic means of x–values.

Estimators of the regression–type that are unbiased have been developed by Mickey (1959) and Williams (1961, 1963), but have not yet been extensively tried. Rao (1969) found Mickey's estimator usually inferior to the classical ratio and regression estimators in natural populations. Singh and Srivastava (1980) proposed a new sampling scheme (SS1, say) for which $\overline{y}_{RG}$ is unbiased. This scheme consists in selecting two units i and j say, with probability proportional to $(x_i - x_j)^2$ and remaining $(n-2)$ units in the sample by SRSWOR. Singh and Srivastava (1980) also proposed another sampling scheme (SS2, say) in which the first unit i say, is select with probability proportional to $\left(x_i - \overline{X}\right)^2$ and remaining $(n-1)$ units by SRSWOR. Employing SS2 they made their proposed regression–type estimator

$$\overline{y}'_{RG} = \frac{n(N-1)}{N(n-1)}\left[\overline{y} - \frac{\sum_{i=1}^{n} y_i\left(x_i - \overline{X}\right)}{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}\left(\overline{x} - \overline{X}\right)\right] \tag{4}$$

completely unbiased. The authors also study the efficiencies of the proposed methods in comparison with standard ratio estimator on samples of size 4, taken from 11 populations of size 20, generated from a bivariate normal population. It turns out that the estimator $\overline{y}_{RG}$ under SS1 is the best in most and next best in all other cases.

3. Unbiased estimators using multi–auxiliary variable

Olkin (1958) generalised the Hartley–Ross unbiased ratio estimator for p–auxiliary variables. The exact expression of the variance for the generalised Hartley–Ross unbiased ratio estimator has been discussed by Ramachandran and Pillai (1976).

Sahoo and Swain (1980) introduced an unbiased ratio–cum–product estimator using two auxiliary variables x_1 and x_2, which was observed to be a particular case of the generalized unbiased estimators due to Williams (1961, 1963) and Mickey (1959). This unbiased estimator is given by

$$\overline{y}_{URP} = \overline{g}_2 + \frac{N-1}{N}\left[\frac{s_{g_1x_1}}{\overline{X}_1} - \frac{s_{g_1x_2}}{\overline{X}_2}\right] \tag{1}$$

where

$$g_{1i} = y_i \frac{\overline{X}_1}{x_{1i}}, \quad g_{2i} = y_i \left(\frac{\overline{X}_1}{x_{1i}}\right)\left(\frac{x_{2i}}{\overline{X}_2}\right) \tag{2}$$

$$\overline{g}_2 = \frac{1}{n}\sum_{i=1}^{n} g_{2i}, \quad s_{g_1x_1} = \frac{1}{n-1}\sum_{i=1}^{n} g_{1i}(x_{1i} - \overline{x}_1), \quad \text{and} \quad s_{g_1x_2} = \frac{1}{n-1}\sum_{i=1}^{n} g_{1i}(x_{2i} - \overline{x}_2) \tag{3}$$

such that $\overline{x}_k$ and $\overline{X}_k$ are respectively the sample and population means of the k–th auxiliary variable ($k = 1, 2$); and x_{ki} is the value of k–th auxiliary variable on the i–th unit.

While constructing $\overline{y}_{URP}$ it was assumed that x_1 is positively and x_2 is negatively correlated with y. In case x_2 is not used, $\overline{y}_{URP}$ reduces to Hartley–Ross unbiased ratio estimator. But if x_1 is not used $\overline{y}_{URP}$ reduces to a Hartley–Ross–type unbiased product estimator based on $\overline{y}'_P$ studied earlier by Gupta and Adhvaryu (1982). It was also shown by Sahoo and Swain (1980) that, the unbiased estimator $\overline{y}_{URP}$ is however not unique and any estimator of the type

$$\left(1 - \frac{\lambda}{\overline{X}_1}\right)\overline{g}_2 + \left(\frac{\lambda}{\overline{X}_1}\right)\overline{g}_1 + \frac{N-1}{N}\left[\frac{s_{g_1x_1}}{\overline{X}_1} - \left(1 - \frac{\lambda}{\overline{X}_1}\right)\frac{s_{g_1x_2}}{\overline{X}_2}\right] \tag{4}$$

is also unbiased for $\overline{Y}$, where λ is any known function of $\overline{X}_1$ and $\overline{X}_2$.

When q auxiliary variables are positively and $(p-q)$ auxiliary variables are negatively correlated with y, Sahoo and Swain (1983) also proposed a Hartley–Ross–type unbiased ratio–cum–product estimator, of the population mean.

3.1. A general class of unbiased estimators

Following Srivastava (1980), Sahoo (1986) proposed a general class of unbiased estimators using p–auxiliary variables. The theoretical background in the construction of the estimator is as follows:

Let $t_i = h(y_i, x_{1i}, \ldots, x_{pi})$ be a function of y_i, x_{1i}, ..., x_{pi} $(i = 1, \ldots, n)$ such that $h(\overline{Y}, \overline{X}_1, \ldots, \overline{X}_p) = \overline{Y}$. The function may contain $\overline{X}_k$ $(k = 1, 2, \ldots, p)$ but independient of $\overline{Y}$. Thus, a class of design biased estimators for $\overline{Y}$ may be defined by

$$\overline{y}_t = \frac{1}{n}\sum_{i=1}^{n} t_i \tag{5}$$

whose bias is given by

$$B = \frac{1}{N}\sum_{i=1}^{n}(t_i - y_i) \tag{6}$$

The estimation of bias involves in expressing B in a simpler form wich further depends on the nature of the function $h(y_i, x_{1i}, \ldots, x_{pi})$. But, sometimes the bias is in the form

$$B = \sum_{k=1}^{p} \theta_k \mathrm{Cov}(f_k, x_k) \tag{7}$$

where f_1, f_2, ..., f_p are functions of y, x_1, ..., x_p; independent of $\overline{Y}$ and θ_1, θ_2, ..., θ_p are known constants. If

$$\widehat{B} = \sum_{k=1}^{p} \theta_k \widehat{\mathrm{Cov}}(f_k, x_k) \tag{8}$$

is an unbiased estimator of B, the class of unbiased estimators of Sahoo (1986) is defined as

$$\overline{y}_U = \overline{y}_t - \widehat{B} \tag{9}$$

The class of estimators represented by $\overline{y}_U$ covers unbiased estimators including those of Hartley and Ross (1954), Gupta and Adhvaryu (1982), Sahoo and Swain (1980, 1983) and Olkin (1958). Sahoo (1986) also pointed out some other interesting unbiased estimators of the class.

References

BEALE, E. M. L. (1962), *Some uses of computers in operational research*, Industrielle Organisation, **31**, 51–52.

DALABEHERA, M. & SAHOO, L. N. (1993), *Unbiased estimators using harmonic mean of the auxiliary variable*, Paper presented at the 47th Annual Conference of the Indian Society of Agricultural Statistics, Tirupati, India.

DESHPANDE, M. N. (1984), *A note on Rao, Hartley and Cochran's method*, J. Indian Soc. Agric. Statist., **36**, 114–116.

GOODMAN, L. A. & HARTLEY, H. O. (1958), *The precision of unbiased ratio–type estimators*, J. Amer. Statist. Assoc., **53**, 491–508.

GUPTA, P. C. & ADHVARYU, D. (1982), *On some unbiased product–type strategies*, J. Indian Soc. Agric. Statist., **34**, 48–54.

HARTLEY, H. O. & ROSS, A. (1954), *Unbiased ratio estimators*, Nature, **174**, 270–271.

IACHAN, R., SINGH, H. P. & UPADHYAYA, L. N. (1987), *On unbiased product estimators*, Gujarat Statist. Rev., **14**, 32–50.

LAHIRI, D. B. (1951), *A method of sample selection providing unbiased ratio estimates*, Bull. Internat. Statist. Rev., **33**(2), 133–140.

MICKEY, M. R. (1959), *Some finite population unbiased ratio and regression estimators*, J. Amer. Statist. Assoc., **54**, 594–612.

MIDZUNO, H. (1952), *On the sampling system with probability proportional to the sum of the sizes*, Ann. Inst. Statist. Math., **3**, 99–108.

MURTHY, M. N. (1964), *Product method of estimation*, Sankhyā, A, **26**, 69–74.

MURTHY, M. N. & NANJAMMA, N. S. (1959), *Almost unbiased ratio estimators based on interpenetrating sub sample estimates*, Sankhyā, **21**, 381–392.

NANJAMMA, N. S., MURTHY, M. N. & SETHI, V. K. (1959), *Some sampling systems providing unbiased ratio estimators*, Sankhyā, **21**, 299–314.

OLKIN, I. (1958), *Multivariate ratio estimation for finite populations*, Biometrika, **45**, 154–165.

PASCUAL, J. N. (1961), *Unbiased ratio estimators in stratified sampling*, J. Amer. Statist. Assoc., **56**, 70–87.

QUENOUILLE, M. H. (1956), *Notes on bias in estimation*, Biometrika, **43**, 353–360.

RAJ, D. (1954), *Ratio estimation in sampling with equal and unequal probabilities*, J. Indian Soc. Agric. Statist., **6**, 127–138.

RAO, J. N. K. (1969), *Ratio and regression estimators. New Developments in Survey Sampling*, N. L. Johnson and H. Smith, Jr. (eds), Wiley, New York, 213–234.

RAO, P. S. R. S. & MUDHOLKAR, G. S. (1967), *Generalised multivariate estimators for the mean of finite populations*, J. Amer. Statist. Assoc., **62**, 1008–1012.

RAO, T. J. (1983), *A new class of unbiased product estimators*, Tech. Report No. 15/83, Indian Statistical Institute, Calcutta.

RAO, T. J. (1987), *On certain unbiased product estimators*, Commun. Statist. Theory Methods, **16**, 963–978.

RAMACHANDRAN, V. & PILLAI, S. S. (1976), *Multivariate unbiased ratio–type estimation for finite sampling*, J. Indian Soc. Agric. Statist., **28**, 71–80.

ROBSON, D. S. (1957), *Application of multivariate polykays to the theory of unbiased ratio–type estimation*, J. Amer. Statist. Assoc., **52**, 511–522.

RUIZ, M. & SANTOS, J. (1989), *Unbiased mean–of–the–ratios estimators*, Statistica, **49**, 617–622.

RUIZ, M. & SANTOS, J. (1990), *Sampling design providing unbiased new product estimators*, Statistica, **50**, 285–288.

SAHOO, L. N. & SWAIN, A. K. P. C. (1980), *Unbiased ratio–cum–product estimator*, Sankhyā, C, **42**, 56–62.

SAHOO, L. N. & SWAIN, A. K. P. C. (1983), *Unbiased ratio–cum–product estimator using multi–auxiliary information*, Gujarat Statist. Rev., **10**, 11–16.

SAHOO, L. N. (1983), *On a method of bias reduction in ratio estimation*, J. Statist. Res., **17**, 1–6.

SAHOO, L.N. (1986), *On a class of unbiased estimators using multi–auxiliary information*, J. Indian Soc. Agric. Statist., **37**, 379–382.

SAHOO, L. N. (1987), *On a class of almost unbiased estimators for population ratio*, Statistics, **18**, 119–121.

SEN, A. R. (1952), *Present status of probability sampling and its use in estimation of farm characteristics*, Econometrica, **20**, 103.

SHAH, D. N. & SHAH, S. M. (1979), *Unbiased product-type estimators*, Gujarat Statist. Rev., **6**, 34–43.

SINGH, M. P. (1967 a), *Multivariate product method of estimation for finite populations*, J. Indian Soc. Agric. Statist., **19**, 1–10.

SINGH, M. P. (1967 b), *Ratio-cum-product method of estimation*, Metrika, **12**, 34–42.

SINGH, H. P. (1989), *A class of unbiased estimators of product of population means*, J. Indian Soc. Agric. Statist., **41**, 113–118.

SINGH, H. P., IACHAN, R. & UPADHAYA, L. N. (1985), *Almost unbiased ratio and product estimators based on interpenetrating sub samples*, Commun. Statist. Theory Methods, **14**, 963–978.

SINGH, P. & SRIVASTAVA, A. K. (1980), *Sampling schemes providing unbiased regression estimators*, Biometrika, **67**, 205–209.

SRIVASTAVA, S. K. (1965), *An estimate of the mean of a finite population using several auxiliary variables*, J. Indian Statist. Assoc., **3**, 189–194.

SRIVASTAVA, S. K. (1980), *A class of estimators using auxiliary information in sample surveys*, Canad. J. Statist., **8**, 253–254.

SRIVASTAVA, V. K. SHUKLA, N. D. & BHATNAGAR, S. (1981), *Unbiased product estimators*, Metrika, **28**, 191–196.

SRIVENKATARAMANA, T. & TRACY, D. S. (1979), *On ratio and product methods of estimation in sampling*, Statist. Neerlandica, **33**, 37–49.

TIN, M. (1965), *Comparison of some ratio estimators*, J. Amer. Statist. Assoc., **60**, 294–307.

VOS, J. W. E. (1980), *Mixing of direct, ratio and product method estimators*, Statist. Neerlandica, **34**, 209-218.

WILLIAMS, W. H. (1961), *Generating unbiased ratio and regression estimators*, Biometrics, **17**, 267–274.

WILLIAMS, W. H. (1963), *The precision of some unbiased regression estimators*, Biometrics, **19**, 352–361.

Rev. Academia de Ciencias. Zaragoza. 49 (1994)

An Alternative Way of Decomposing Stimulus Variability in Confirmatory MDS.

Vera, J. F., García, P. A., González, A.

Departamento de Estadística e Investigación Operativa

Facultad de Ciencias, Universidad de Granada, 18071 Granada

Abstract

We propose an alternative model designed for estimating the variability associated with stimuli within a metric Multidimensional Scaling (MDS) analysis of dissimilarity data deriving from one subject and based on the confirmatory model proposed by Ramsay. The variability factors are estimated by maximum likelihood using the lognormal distribution as the dissimilarity data distribution. The choice between the Ramsay model and ours is discussed.

1. Introduction.

The first confirmatory MDS model (see Schiffman *et al.* [7]) was that proposed by Ramsay [3], which offered the possibility of analyzing data from the point of view of statistical deduction. The first subsequent work in this context, showing the necessity of variability decomposition in factors associated with each element of a model was also conducted by Ramsay [4]. He examined one variability factor dependent upon the individual being questioned and another dependent upon each pair of stimuli on which the individual bases his judgement. At the same time he proposed considering separately the variability associated with the stimuli by using an additive decomposition model.

Estimating these components through maximum likelihood in the MDS model causes various inconveniences in the interpretation of the results and requires a great deal of computation, which is determined by the estimation of an additive model within the lognormal distribution. In many cases this estimation has to be calculated after the estimation of the remaining parameters of the model, since otherwise the computations become much more numerous and complex.

The need for variability decomposition had to be proven by experience as the distances estimated by the MDS models do not adjust themselves perfectly to the observed dissimilarity values. There are many very diverse factors that cause the residuals to have a value

other than zero. There are two direct ways of controlling the variation in the residuals, or errors: firstly, by taking into account the different variation components, thus making a great refinement of the analysis possible by using more precise estimations; and secondly, by a summary of the variation components, which may also add many important aspects to the data interpretation.

One aspect worth considering is that of explicitly establishing the possibility that the value of each stimulus might be independent of the global value of the variance for each pair, γ_{ij}. This is achieved by the decomposition of γ_{ij} into specific components for each stimulus, which will be α_i. Therefore these components will be interpreted as the relative contributions to the variance for each stimulus in each pair and the data interpretation is conditioned by the stimulus variability factors. Thus, the choice of the procedure that establishes the relationship between the variation components constitutes an important element to bear in mind inside the analysis and, as will be shown below, different decomposition models may result in different interpretations.

The subject's perception of a particular stimulus, i, may be variable or undetermined, causing the classifications of pairs to force the stimuli to have a high degree of variability. This may be reflected in the fact that the value belonging to the component which corresponds to a defined stimulus, α_i, is close to the total variability of each pair, γ_{ij}. On the other hand, a certain stimulus might act as a typical stimulus or one used as a reference to those with which it is compared, thus causing a minor variability in those judgements in which this stimulus is involved and identified by a value of its variability component, α_i, which will be lower than the rest of the stimuli with which it is compared.

All this requires that we should take special care with the decomposition model employed, so that the interpretation of the variability designated to the stimulus should not lead to contradictory conclusions in the data analysis. The decomposition model proposed in this paper explains the influence exerted by each stimulus on the individual being questioned, when the distribution model is lognormal.

2. Description of the Model.

Following Ramsay's [5] model notation, we use n stimuli, one subject and T replications. Generally, the differing data will comprise T squared matrices $n \times n$ from dissimilarity data between each pair of stimuli, the elements of which will be represented by d_{ijt}, with $i, j = 1, \ldots, n$ associated with each pair of stimuli (i, j) and $t = 1, \ldots, T$ and for each response t. The configuration matrix to be estimated will be represented by X and the point corresponding to each stimulus, i, will be represented by x_i.

The distance model will be the Euclidean one, where the distance between each pair of

points i and j, associated to their corresponding stimuli, will be indicated by d^*_{ij}, expressed as follows

$$d^*_{ij} = \left[\sum_{m=1}^{k}(x_{im} - x_{jm})^2\right]^{1/2} \quad (1)$$

In this situation, from the statistical point of view, the data are considered as being values taken from the corresponding random variable D_{ij}, each being independent of the other. They formally take on values in $\mathbb{R}^+$ and are distributed around the corresponding central value $log(d^*_{ij})$, according to a two-parameter lognormal distribution represented as

$$D_{ij} \sim \Lambda(\log(d^*_{ij}), \sigma^2_{ij}) \quad (2)$$

where the variance σ^2_{ij} is considered to be dependent upon each pair (i, j) or constant with the value σ^2.

Although obtaining the configuration matrix associated with the dissimilarity data constitutes the central aspect of MDS, one of the additional advantages of the confirmatory model lies in the possibility of estimating the variability factors that influence the obtention of this configuration.

The variability model proposed in this paper (see Vera [10]) is broken down in the following manner:

$$\gamma^2_{ij} = (\alpha^2_i \alpha^2_j)^{1/2} \quad (3)$$

The variability generated by the stimuli in the study and analysed by means of this decomposition offers important advantages when used in a lognormal context. Although the additive decomposition model allows for an easy interpretation of the estimated values when the distribution chosen for the data representation is the normal distribution, the additive model joined to the lognormal distribution creates several disadvantages that make the interpretation of the results of the analysis more difficult.

The use of the multiplicative model in the lognormal case solves these problems since its computational cost is relatively low and the interpretations of the estimated values are similar to those which are found when an additive model is used jointly with the normal distribution. Let us see the estimation of these components.

The loglikelihood associated to the model will remain as

$$\log L = -\frac{1}{2}\sum_{i \neq j}\left[\frac{T}{2}\log(\alpha^2_i \alpha^2_j) + \frac{1}{(\alpha^2_i \alpha^2_j)^{1/2}} S_{ij}\right] - \sum_{i \neq j}\sum_{t}\log(d_{ijt}) - M\log(\sqrt{2\pi}) \quad (4)$$

where

$$S_{ij} = \sum_t \log^2 \left(\frac{d_{ijt}}{d^*_{ij}} \right) \tag{5}$$

and M is the number of dissimilarity data actually observed.

The estimation by maximum likelihood of the variability factors is carried out by imposing the following constraints

$$\sum_i \alpha_i^2 = n \tag{6}$$

These constraints lead to the use of a penalty function (see Fiacco & McCormick [1]). The term of this function $Q(X, \alpha) = q(X) + q(\alpha)$ associated with the restrictions of the subject's typical errors, $q(\alpha)$, is expressed by,

$$q(\alpha) = -\frac{1}{2} \left(\sum_i \alpha_i^2 - n \right)^2 \tag{7}$$

Taking the partial derivative of the $\log L$ with respect to α_p^2 the following is obtained:

$$\frac{\partial \log L}{\partial \alpha_p^2} = \sum_{i \neq j} \left[\left(\frac{T}{(\alpha_i^2 \alpha_j^2)^{1/2}} - \frac{S_{ij}}{(\alpha_i^2 \alpha_j^2)} \right) \left(\frac{\partial}{\partial \alpha_p^2} (\alpha_i^2 \alpha_j^2)^{1/2} \right) \right] = 0 \tag{8}$$

Simplifying the previous expression gives these results

$$2Rn(\alpha_p^2)^{1/2} = \sum_j \frac{S_{pj} + S_{jp}}{(\alpha_j^2)^{1/2}} \tag{9}$$

and finally

$$\alpha_p^2 = \left[\frac{1}{2Tn} \sum_j \frac{S_{pj} + S_{jp}}{\mid \alpha_j \mid} \right]^2 \tag{10}$$

where the distances, d^*_{ij}, are given in terms of the configuration matrix, X, estimated by the implicit equation of Ramsay [3]

$$x_{pq} \sum_j t_{pj} = \sum_j x_{jq} t_{pj} \tag{11}$$

$$p = 1, \ldots, n. \qquad q = 1, \ldots, K$$

and where

$$t_{pj} = \frac{1}{d^*_{pj}} \sum_t \left[\log \left(\frac{d_{pjt}}{d^*_{pj}} \right) + \log \left(\frac{d_{jpt}}{d^*_{pj}} \right) \right] \tag{12}$$

Table 1.—Variability of recreative activities.

STIMULI	ADITIVE MODEL	MULTIPLICATIVE MODEL
1 CONCERT	0.50	0.2528605
2 MUSEUM	2.07	1.8750590
3 THEATRE	1.30	0.3061745
4 CINEMA	2.54	0.8480011
5 TELEVISION	2.64	4.8404140
6 CONFERENCE	0.00	0.1742502
7 READING	1.35	0.7342721
8 HOCKEY	0.43	0.1304034
9 BALLET	0.00	0.0026923
10 DEBATE	1.09	0.7239251
11 FASHION	0.28	1.9462000
12 DOC-CINEMA	0.00	0.0890419
13 EXHIBITION	0.67	0.3169389
14 SHOPPING	1.68	2.5689700
15 RESTAURANT	1.24	0.1907957

Our experience with this structural hypothesis has been that much less calculation time is needed for the total computation of these estimators of the model. Furthermore, the choice of a geometric approach to the variability decomposition is more suited to the structure of the variance in the lognormal models.

3. An Illustrative Example.

First of all it is necessary to emphasize that the resolution of the implicit equations which determine the obtaining of the estimators that maximize the likelihood is carried out numerically, since it is not possible to obtain estimators explicitly. To do this we have used a process similar to that described by Takane *et al.* [8], which consists of a cycle of main iterations, inside each of which the parameters are updated in blocks. To update each block of parameters, a new cycle of secondary iterations is employed in which this updating of the parameters is achieved by means of a procedure of conjugate gradient containing an algorithm that relies upon cubic interpolation and extrapolation, has been used, (see Fletcher [2]) to determine the optimal step size, Torgerson's [9] algorithm being used to obtain the initial configuration matrix.

Table 2.—The order of stimuli according to their variability (least to greatest).

ADDITIVE MODEL		MULTIPLICATIVE MODEL	
0.00 BALLET	(9)	0.00 BALLET	(9)
0.00 CONFERENCE	(6)	0.09 DOC-CINEMA	(12)
0.00 DOC-CINEMA	(12)	0.13 HOCKEY	(8)
0.28 FASHION	(11)	0.17 CONFERENCE	(6)
0.43 HOCKEY	(8)	0.19 RESTAURANT	(15)
0.50 CONCERT	(1)	0.25 CONCERT	(1)
0.67 EXHIBITION	(13)	0.31 THEATRE	(3)
1.09 DEBATE	(10)	0.32 EXHIBITION	(13)
1.24 RESTAURANT	(15)	0.72 DEBATE	(10)
1.30 THEATRE	(3)	0.73 READING	(7)
1.35 READING	(7)	0.85 CINEMA	(4)
1.68 SHOPPING	(14)	1.88 MUSEUM	(2)
2.07 MUSEUM	(2)	1.95 FASHION	(11)
2.54 CINEMA	(4)	2.57 SHOPPING	(14)
2.64 TELEVISION	(5)	4.84 TELEVISION	(5)

Two convergence criteria have been simultaneously employed in our model: firstly, a purely geometric one (the gradient method), in which not only the direction is controlled, but also the module of the gradient vector to determine the end of the iterative process; and secondly, a statistical procedure based on the χ^2 contrast, in view of the fact that the method does have a strong statistical character.

To illustrate the method, we examine an example from Ramsay [6] in which he analysed 105 values of the dissimilarity given by a particular subject about 15 recreative activities: concert, museum, theatre, cinema, television, conference, reading, hockey, ballet, debate, fashion, documentary cinema, exhibition, shopping and restaurant.

The results are shown in Table 1, where the variability associated with each stimulus is compared to the multiplicative decomposition model explained earlier and the additive proposal of Ramsay [4]. In Table 2 the stimuli are put in order from the least variability to the greatest.

These results show the differences that exist upon considering a multiplicative model instead of an additive one. The results shown in Table 2 show clearly the effect that the multiplicative model has on the distribution of the variabilities, assigning extreme values only to those stimuli which exert an extreme influence and distributing the remaining

variability among the stimuli which have a moderate impact on the subject. This is determined by the same degree of extreme variability which both a very familiar stimulus or a completely unknown or rejected one may produce on the subject's answers. The rest of the stimuli, which cause a moderate response in the subject tend to produce moderate variability values.

In this example it can be seen how stimuli with opposite variabilities such as *ballet* and *television* and, generally speaking, the groups of stimuli which have greater or lesser variability are distinguished in the same way by both models. Nevertheless, stimuli to which the additive model assigns no variability, such as *conference* and *doc cinema* are distinguished by the multiplicative model, whilst activities such as *concert* and *theatre* or *fashion* and *shopping*, which produce very pronounced differences between each other in the additive model, are classified as similar influences when the multiplicative model is applied.

Even though the criteria for choosing between the two models is based upon the researcher's own opinion, this example clearly shows the influence of the choice of the decomposition used in the final interpretation of the analysed data. Therefore, the resulting individual's profile, and consequently the final interpretation of the configuration matrix, changes considerably depending upon whether the additive model or the multiplicative model is used, and although the estimation of the configuration matrix is not altered by the model employed, the choice between an additive and a multiplicative model must be considered in any analysis.

References

[1] Fiacco, A. V. & McCormick, G. P. (1968).- Nonlinear Programming: Sequential Unconstrained Minimization Techniques. *John Wiley and Sons, Inc.*

[2] Fletcher, R. (1980).- Practical Methods of Optimization Vol.1. Unconstrained Optimization. *John Wiley and Sons.*

[3] Ramsay, J. O. (1977).- Maximum Likelihood Estimation in Multidimensional Scaling. *Psychometrika*, 42, 241-266.

[4] Ramsay, J. O. (1978).- Confidence Regions for MDS Analysis. *Psychometrika*, 43, 145-160.

[5] Ramsay, J. O. (1982).- Some Statistical Approaches to MDS Data. *J. R. Statist. Soc. A*, 145, 285-312.

[6] Ramsay, J. O. (1986).- MULTISCALE II MANUAL. *Department of Psychology. McGill University.*

[7] Schiffman, S. S., Reynolds, M. L. & Young, F. W. (1981).- Introduction to Multidimensional Scaling. Theory, Methods and Applications. *Academic Press, Inc.*

[8] Takane, Y., Young, F. W. & de Leew, J.(1977).- Nonmetric Individual Differences MDS: An alternating Least Squares Method with Optimal Scaling Features. *Psychometrika*, 42, 7-67.

[9] Torgerson, W. (1958).- Theory and Methods of Scaling. *Wiley, New York.*

[10] Vera, J. F. (1993).- Inference in MDS and its Computational Treatment. *Unpublished Doctoral dissertation. University of Granada.*

Rev. Academia de Ciencias. Zaragoza. 49 (1994)

La dinámica de sistemas como técnica metodológica en la educación ambiental.

J.A. Sánchez (1) y M. Villagrasa (2)

(1) Deptº de Ciencias de la Tierra
Universidad de Zaragoza

(2) Centro de Recursos de Cariñena. M.E.C.

Abstract.

We exemplify different applications of the dynamic systems use in order to make up usable models on environment education: radioactive decreasing, increasing of the population without control, relation between prey and predator, reversible quimical reactions, with or without a losing of reagent material....

They are exemplified in this way, the numerous possibilities that dynamic technics of systems can be use full to the educational development.

1. Introducción.

La palabra sistema es de gran actualidad en la ciencia, se refiere a un todo o conjunto en el que se pueden distinguir diversos elementos que actúan unos sobre otros, o se influyen mutuamente de algún modo, de manera que las propiedades del sistema no son coincidentes con la suma de los elementos que lo conforman.

En el estudio de los sistemas interesa más el conocimiento de las relaciones entre los elementos interactuantes que la naturaleza exacta de estos elementos, y ésto es los que hace la ecología cuando estudia los ecosistemas, donde básicamente considera las entradas y salidas de

materia y energía, la organización e interacciones entre los componentes, así como el cambio y evolución que se experimenta.

Toda la cubierta viva de la Tierra se puede decir que constituye un ecosistema - el mayor que existe - que recibe el nombre de biosfera. Pero también se denominan ecosistemas a cualquier parte más pequeña de esa biosfera (un lago, un bosque, el océano, etc.). El conocimiento del medio ambiente pasa por la comprensión de los distintos ecosistemas y de su interacción con la actividad antrópica.

Con la dinámica de sistemas (DS) se aporta una herramienta capaz de simular las interacciones y cambios que se producen en un ecosistema a partir básicamente del conocimiento de su estructura. El ordenador es el medio que hace operativa la dinámica de sistemas.

2. La dinámica de sistemas.

La dinámica de sistemas (DS) nace como una técnica que permite analizar los sistemas y simular sus comportamientos en el tiempo. J.W. Forrester, ingeniero de sistemas del Instituto de Tecnología de Masschusets (MIT), desarrolló este método durante la década de los años cincuenta. La primera aplicación fue el análisis de la estructura de una compañía eléctrica Norteamericana y el estudio de las oscilaciones de ventas de dicha empresa. Posteriormente, la DS se aplicó al estudio de sistemas mecánicos ecológicos (especialmente poblaciones), al comportamiento de sistemas sociales y a la planificación urbana, pero la DS se popularizó con su utilización por Forrester para la elaboración del "World Dynamics" ó "Modelo del mundo", que fue básico para la elaboración del primer Informe del Club de Roma, donde aparecen por primera vez los términos, actualmente en uso, de "crecimiento cero", "desarrollo sostenido", "límites al crecimiento", etc.

La difusión y bajo precio de los ordenadores y la disponibilidad de software de fácil utilización (Stella, Dynamo, etc.) hacen de la DS una técnica de gran futuro a nivel técnico, científico y educativo.

Básicamente la DS es una técnica que permite:

A) Establecer la estructura del sistema, determinando que elementos son más significativos y como están relacionados.

B) Simular la evolución temporal de los elementos del sistema, según las circunstancias en que se desenvuelva el sistema, es decir, su funcionamiento.

2.1.- Componentes de un sistema.

Un sistema presenta los siguientes componentes estructurales:

a) Elementos: Son los componentes fundamentales del sistema y cada elemento es la representación simplificada de alguna característica de la realidad objeto de estudio.

b) Relaciones entre los elementos: En un sistema los elementos están interrelacionados por redes de comunicación que aportan materia, energía, información, etc.

c) Límites: Los sistemas tienen espacio donde se encuentran sus elementos. Pueden existir elementos externos al sistema (elementos exógenos) que no actúan directamente sobre el sistema, pero sí sobre algún elemento interno (endógeno) y que, por tanto, deben ser considerados.

Entre los componentes estructurales se establecen las siguientes relaciones funcionales:

α) Flujos de materiales o de información o de energía, que circulan entre variables de estado. La circulación se efectúa a través de las redes de comunicación.

β) Mecanismos de control de los flujos (válvulas, grifos)

γ) Retardos, que resultan de las discrepancias entre unidades de tiempo y velocidades de circulación de los flujos.

δ) Bucles de ralimentación (feedback) o cadenas de causalidad o influencias circulares entre elementos.

La representación formal de un sistema es un modelo matemático, y con la DS se construye este modelo de un modo sencillo, mediante la elaboración de diagramas de causalidad y su expresión en una simbología muy didáctica debida a Forrester, que con pequeñas modificaciones han adoptado los distintos softwares de DS.

2.2.- Diagramas causales y de Forrester.

Un diagrama causal es una representación gráfica en la que aparecen formalizados los elementos del sistema y se establecen las relaciones que existen entre ellos, haciendo constar cual es el signo de

variación esperado entre cada par de elementos. Así, cuando se indica la relación:

a (presa) ——————→ + b (+ predador)

se quiere significar que las variables a y b se mueven en el tiempo en el mismo sentido (cuando crece a, crece b).

Evidentemente, una relación del tipo:

a (predador) ——————→ - c (- presa)

significa todo lo contrario; es decir que, las variables a y c se mueven a lo largo del tiempo en sentido contrario: cuando una crece la otra decrece.

Especial interés tienen los bucles de realimentación que en un diagrama causal se indican por medio de una "flecha" que partiendo de una variable "vuelve" a ella, después de un camino más o menos largo.

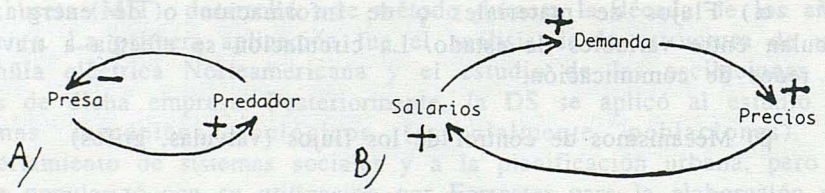


Fig. 1.- Diagramas causales: A) Diagrama causal estable.
B) Diagrama causal explosivo o autorreforzado.

Así, la Fig. 1 A) muestra un diagrama causal estable: el número de individuos de la especie depredada (presa), actúa positivamente sobre el número de individuos de la especie depredadora (predador), simultáneamente, el número de individuos depredadores actúa negativamente sobre el número de individuos depredados. Otros bucles como el de la Fig. 1 B), son explosivos o autorreforzados, como sucede en la espiral inflacionista.

A partir de los diagramas causales se pasa a los diagramas de Forrester, cuya simbología se expresa en la Fig. 2. Con esta simbologia los programas informáticos DS construyen las ecuaciones en función de los esquemas de relación que nosotros realizamos gráficamente. Tan sólo es preciso, en todos los casos, asignar valor a las constantes, valores iniciales a las variables de nivel, tiempo de simulación y valor del incremento de tiempo considerado. En determinados casos los datos se pueden introducir en forma de gráficos o tablas.

Los resultados, también denominados trayectorias, corresponden a la evolución temporal de las variables y aparecen expresados, tanto gráfica como numéricamente. Las modificaciones en los datos iniciales son muy fáciles de realizar, pudiendo así simular diferentes escenarios.

Nube: Representa una fuente o un sumidero; puede interpretarse como un nivel que no tiene interés por ser prácticamente inagotable.

Nivel: Representa una acumulación de flujo, es una variable de estado.

Flujo: Variación de nivel, representa un cambio en el estado del sistema.

Canal de material: Canal de transmisión de una magnitud física que se conserva.

Canal de información: Canal de transmisión de una cierta información, que no es necesario que se conserve.

Variable auxiliar: Una cantidad con un cierto significado físico en el mundo real y con un tiempo de respuesta instantáneo.

Constante: Un elemento del modelo que no cambia de valor.

Retardo: Un elemento que simula retrasos en la transmisión de información o material.

Variable exógena: Variable cuya evolución es independiente de la del resto del sistema. Representa una acción del medio sobre el sistema.

Fig. 2.- Simbología de la dinámica de sistemas, introducida por Forrester

3. Ejemplo de aplicación de la DS.

A continuación se muestran algunos ejemplos de aplicación de la DS mediante el programa Stella, que pretenden ser una muestra de las enormes posibilidades que esta herramienta tiene.

3.1.- Crecimiento y decrecimiento explosivos.

Son dos sencillos ejemplos de como, en función de una tasa, aumenta o disminuye exponencialmente el valor de una variable de estado ó nivel (Fig. 3). En la naturaleza se tienen numerosos ejemplos de esta evolución en el tiempo: desintegración radiactiva, crecimiento de plagas o epidemias sin control, agotamiento de manantiales, etc.

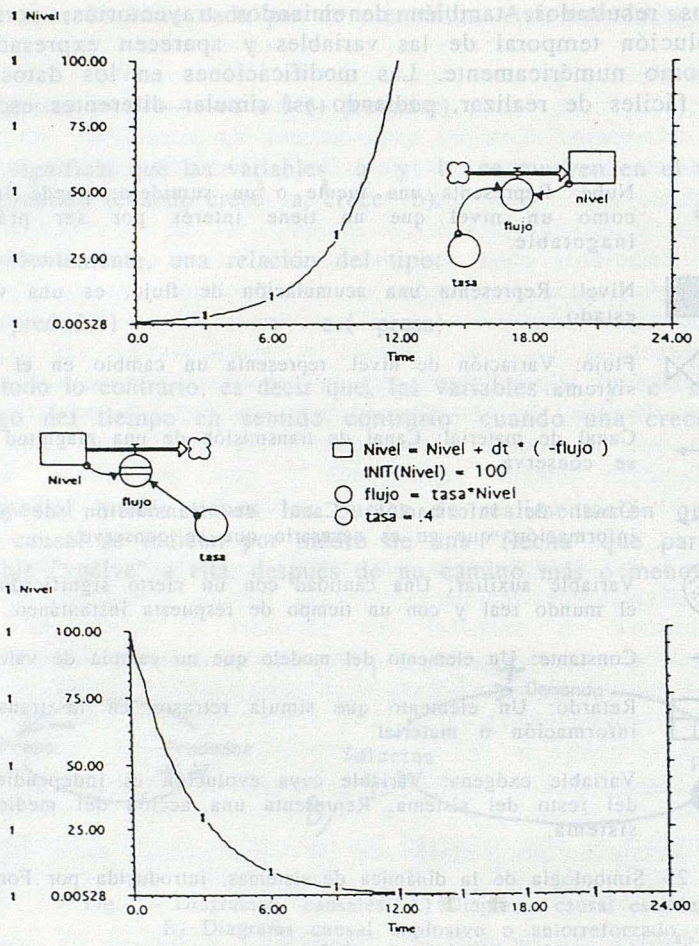
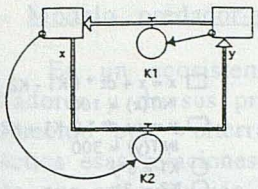
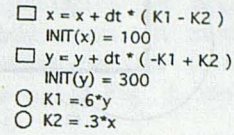


Fig. 3.- Diagramas DS, ecuaciones y trayectoria de la variable para casos de crecimiento o decrecimiento explosivos o autorreforzados.

3.2.- Flujos reversibles, conservativos y no conservativos.

Se establece un flujo de material (o energía) entre dos variables de estado, controladas por dos tasas distintas. En función de esas tasas, los valores iniciales de las variables evolucionan hacia un equilibrio dinámico (conservativo). Un ejemplo podría ser una reacción química reversible (fig. 4 a). Si en una de las variables existe un flujo que extrae material (o energía), la evolución de las variables es claramente distinta (fig. 4 b). Un ejemplo de esta evolución podría ser una reacción química en que parte del material resulta precipitado, saliendo del sistema.





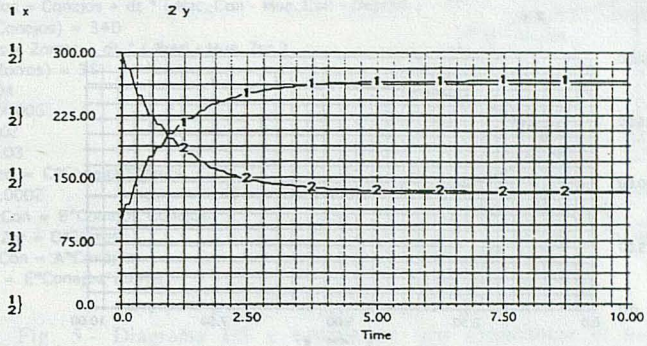
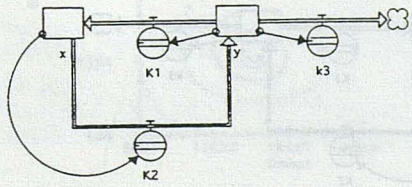


FIG.4a-Diagramas DS, ecuaciones y trayectorias de las variables en una reacción reversible conservativa.



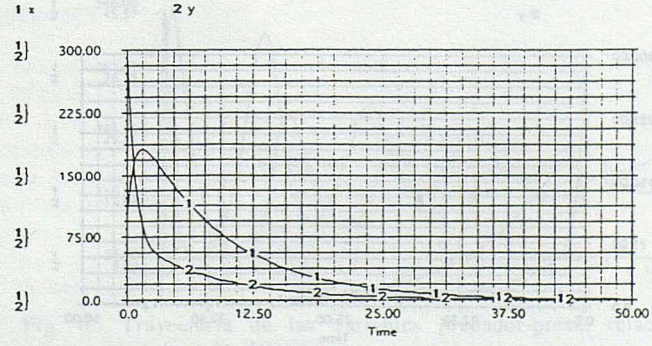


FIG.4b-Diagramas DS, ecuaciones y trayectorias de las variables en una reacción reversible no conservativa.

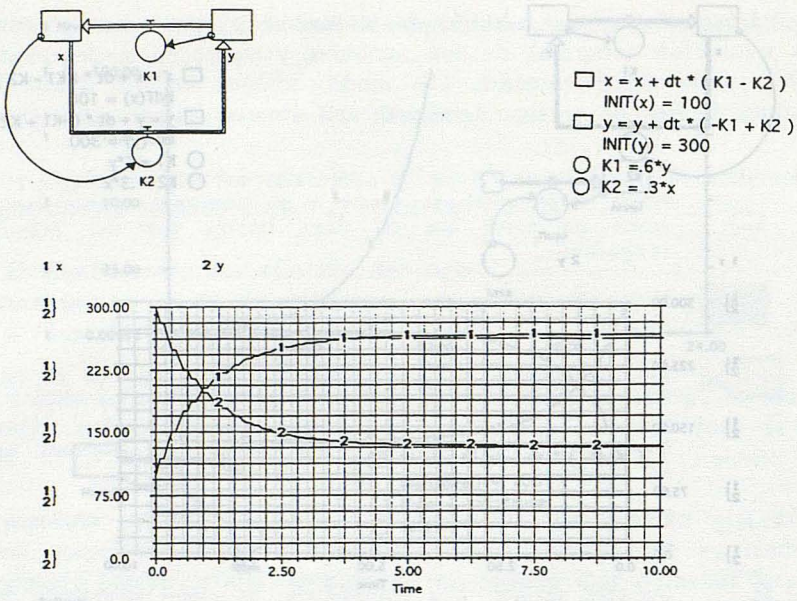


FIG.4a-Diagramas DS, ecuaciones y trayectorias de las variables en una reacción reversible conservativa.

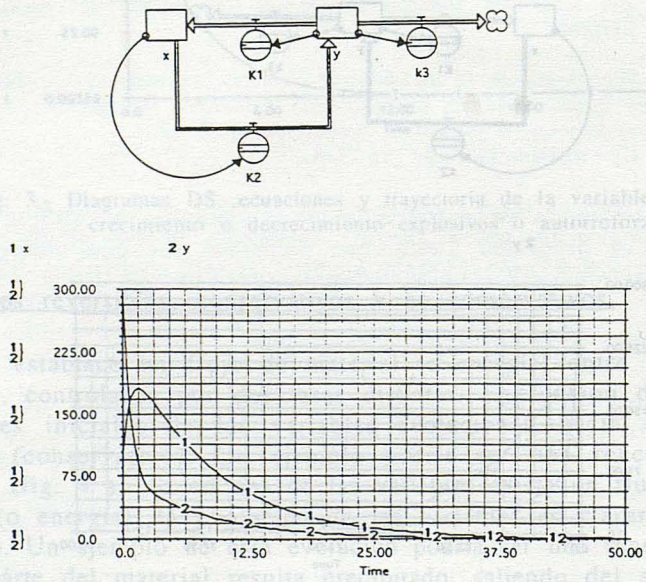


FIG.4b-Diagramas DS, ecuaciones y trayectorias de las variables en una reacción reversible no conservativa.

3.3.- Modelo predador-presa (modelo Volterra).

En un ecosistema la evolución de la población de organismos predadores y de sus presas está intimamente relacionada. El modelo fue establecido por Volterra, y la DS permite mostrar de una forma muy didáctica esas relaciones, tal y como puede verse en la Fig. 5, del mismo modo que su evolución se representa en la Fig. 6.

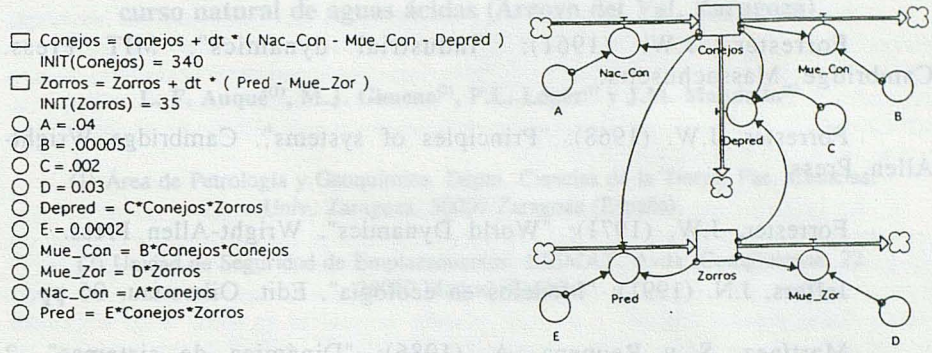


Fig. 5.- Diagrama DS y ecuaciones que representan el modelo Volterra.

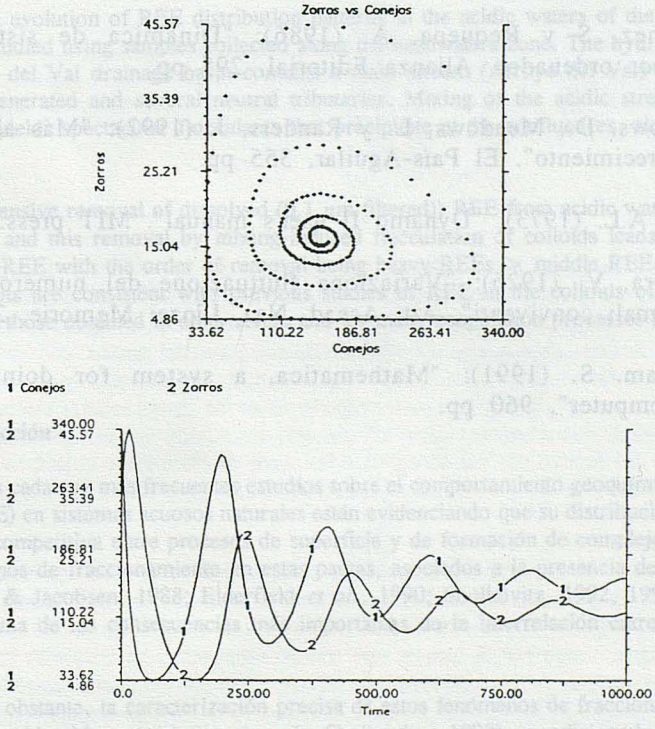


Fig. 6.- Trayectoria de las variables predador-presa. relacionadas entre sí, y en función del tiempo.

4. Bibliografía.

Aracil, J. (1992): "Introducción a la dinámica de sistemas". Alianza Editorial Textos. 398 pp.

Barry, R. (1987): "Stella for business". Hannover, N.H., High Performance Systems.

Forrester, J.W. (1961): "Industrial dynamics". MIT Press. Cambridge Massachusetts.

Forrester, J.W. (1968): "Principles of systems". Cambridge Wright-Allen Press.

Forrester, J.W. (1971): "World Dynamics". Wright-Allen Press.

Jeffers, J.N. (1991): "Modelos en ecología". Edit. Oikos-tau. 96 pp.

Martínez, S y Requena, A. (1986): "Dinámica de sistemas". 2 Modelos. Alianza Editorial, 295 pp.

Martínez, S y Requena, A. (1986): "Dinámica de sistemas". 1 Simulación por ordenador. Alianza Editorial, 295 pp.

Meadows, D., Meadows, L. y Randers, J. (1992): "Más allá de los límites del crecimiento". El País-Aguilar, 355 pp.

Pugh, A.L. (1973): "Dynamo II user's manual". MIT press.

Volterra, V. (1926): "Variazione fluttuazione del numero díndivui in specie animali conviventi". Att. Accad. Naz. Linzei Memorie, 2, 31-113.

Wolfram, S. (1991): "Mathematica, a system for doing mathematica by computer", 960 pp.

Rev. Academia de Ciencias. Zaragoza. 49 (1994)

Pautas de evolución en la distribución de Tierras Raras a lo largo de un curso natural de aguas ácidas (Arroyo del Val, Zaragoza).

L. F. Auqué[1], M.J. Gimeno[2], P.L. López[1] y J.M. Mandado[1]

(1) Área de Petrología y Geoquímica. Depto. Ciencias de la Tierra. Fac. Ciencias. Univ. Zaragoza. 50009 Zaragoza (España).

(2) Unidad de Seguridad de Emplazamientos. CIEMAT. Avda. Complutense, 22. 28040 Madrid (España).

Abstract

The evolution of REE distribution patterns in the acidic waters of the Arroyo del Val stream is studied using samples collected along the headwaters zone. The hydrologic system of the Arroyo del Val drainage basin contains a main stream (Arroyo del Val) of acidic waters naturally generated and several neutral tributaries. Mixing of the acidic stream with neutral tributaries yields spectacular flocculants that precipitate at the confluences, along the principal stream.

Extensive removal of dissolved (0.1 μm filtered)) REE from acidic waters occurs along flow path; and this removal by mixing-induced flocculation of colloids leads to fractionation among the REE with the order of removal being heavy REEs > middle REEs > light REEs. These results are consistent with previous studies of REE in the colloids of this system but opposite to those obtained in other rivers and for many coagulation processes in estuaries.

1. Introducción

Los cada vez más frecuentes estudios sobre el comportamiento geoquímico de las Tierras Raras (REE) en sistemas acuosos naturales están evidenciando que su distribución depende de la actuación competitiva entre procesos de superficie y de formación de complejos. La existencia de fenómenos de fraccionamiento en estas pautas, asociados a la presencia de fases coloidales (Goldstein & Jacobsen, 1988; Elderfield *et al.*, 1990; Sholkovitz, 1992, 1993; etc.), parece constituir una de las consecuencias más importantes de la interrelación entre ambos tipos de procesos.

No obstante, la caracterización precisa de estos fenómenos de fraccionamiento presenta todavía considerables problemas (p. ej. Sholkovitz, 1992), condicionados por las bajas concentraciones en las que se encuentran estos elementos y por las dificultades en separar la fracción asociada a las fases coloidales.

El sistema geoquímico del Arroyo del Val en la provincia de Zaragoza (Gimeno, 1991; Gimeno *et al.*, 1994; Auqué *et al.*, 1993) está constituido por un curso principal de aguas ácidas (Arroyo del Val), generadas por el lavado de los materiales paleozoicos que atraviesa (con abundantes sulfuros metálicos dispersos), afectado por sucesivos aportes de afluentes neutro-básicos que neutralizan progresivamente las aguas del arroyo principal. En los puntos de confluencia de estos afluentes con el Arroyo del Val se producen procesos de mezcla con una floculación masiva de coloides blanquecinos que tapizan el fondo del arroyo.

En este sistema se conjugan varias circunstancias que lo hacen especialmente interesante para el estudio del comportamiento de las REE: se trata de un sistema inicialmente ácido, de baja temperatura y, a priori, capaz de transportar y/o movilizar una importante concentración de estos elementos; la presencia de sucesivos procesos de floculación coloidal facilita el análisis de su influencia en la distribución de Tierras Raras; y la abundancia de coloides en algunos tramos permite un muestreo relativamente fácil.

La idoneidad de este conjunto de características ya ha sido verificada en estudios preliminares. A partir de los datos obtenidos en el primero de los puntos de floculación del Arroyo del Val (figura 1), Auqué *et al.* (1993) indican la existencia, tanto para las aguas ácidas como para las fases coloidales, de una pauta de REE definida por un enriquecimiento neto en las Tierras Raras pesadas (HREE) y con una marcada convexidad en torno a los contenidos normalizados de las Tierras Raras intermedias (IREE). Y señalan la presencia de un fraccionamiento en el proceso de floculación, caracterizado por un enriquecimiento sistemático en las Tierras Raras más pesadas.

En este artículo se exponen los resultados de un muestreo más completo, analizándose la evolución de la distribución de REE en las aguas ácidas a lo largo del curso superior del Arroyo del Val y verificando las pautas de fraccionamiento ligadas a los sucesivos fenómenos de floculación coloidal.

2. Metodología

El muestreo realizado para el análisis de Tierras Raras, se localiza en el tercio superior del curso del Arroyo del Val y se inscribe dentro de otro más amplio, en el que fueron tomadas muestras para la determinación de elementos mayores y menores de las aguas de este sistema. En la figura 1 se muestra un esquema simplificado de esta parte del arroyo, con los principales puntos de mezcla y floculación coloidal, tras la confluencia con afluentes (primero, segundo, cuarto y quinto puntos de mezcla) y aportes hipodérmicos de aguas neutras (tercer punto de mezcla). A lo largo de este tramo se realizó un completo desmuestre que incluyó muestras de las soluciones ácidas del Arroyo del Val, tanto entre las zonas de floculación coloidal como en los propios puntos de floculación, y muestras de los afluentes de aguas neutras que desembocan en este tramo.

Los procesos de floculación coloidal, inducidos por la mezcla de soluciones, se distribuyen de forma perfectamente localizada tras la confluencia de los afluentes de aguas neutras. Pero en esos puntos los efectos de esos fenómenos de floculación varían tanto espacial como temporalmente, razón por la que en este trabajo únicamente se utilizarán las muestras tomadas antes y/o después de los fenómenos de mezcla y floculación (cuya situación se indica en la figura 1), excluyendo las tomadas durante la actuación de esos procesos. De esta forma, se considerarán los datos correspondientes a las soluciones acuosas una vez que el conjunto de procesos actuantes han homogeneizado sus efectos sobre la evolución de las REE, evitándose las

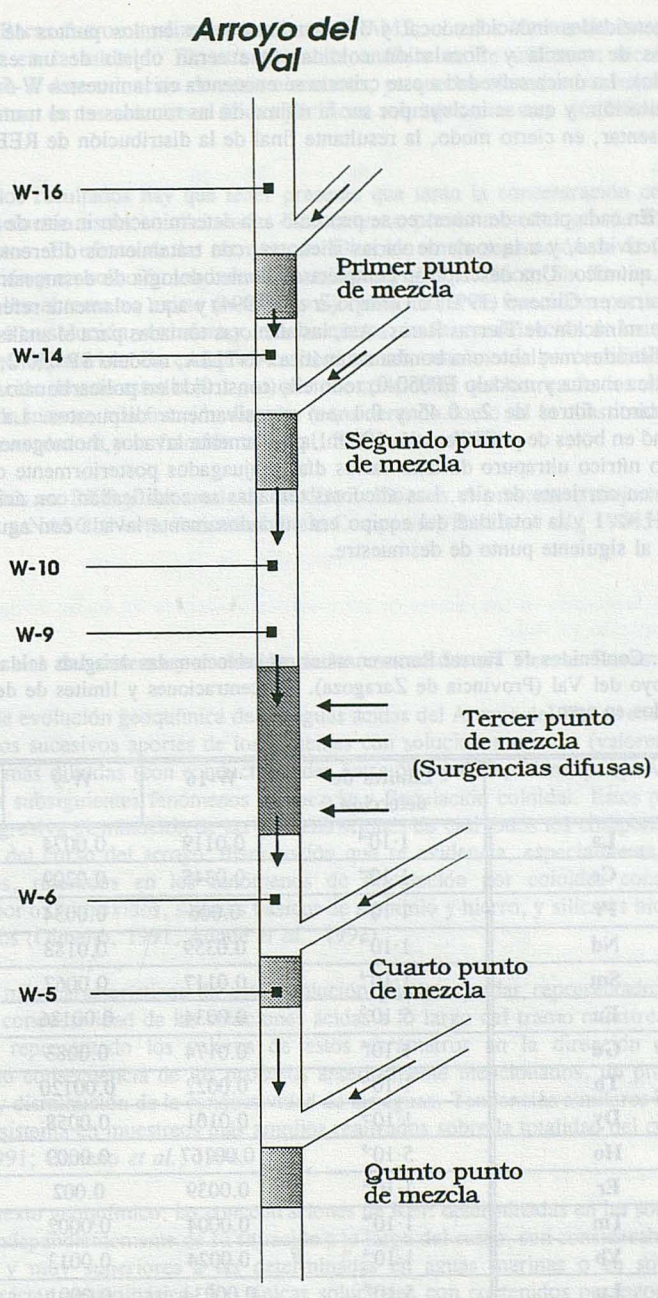


Figura 1. Esquema de la cabecera del Arroyo del Val y localización del muestreo. Los puntos de floculación coloidal (asociados a procesos de mezcla con soluciones neutras) se indican con áreas rellenas a lo largo del tramo estudiado.

heterogeneidades inducidas local y/o instantáneamente en los puntos de confluencia por los procesos de mezcla y floculación coloidal (que serán objeto de un estudio posterior más detallado). La única salvedad a este criterio se encuentra en la muestra W-5, situada en un punto de floculación, y que se incluye por ser la última de las tomadas en el tramo muestreado (fig.1) y representar, en cierto modo, la resultante final de la distribución de REE en la evolución del sistema.

En cada punto de muestreo se procedió a la determinación in situ de pH, Eh, temperatura y conductividad, y a la toma de varias alicuotas, con tratamientos diferentes, para su posterior análisis químico. Una descripción completa de la metodología de desmuestre y análisis utilizada puede verse en Gimeno (1991) o Gimeno *et al.* (1994) y aquí solamente referiremos la que afecta a la determinación de Tierras Raras. Así, las alícuotas tomadas para el análisis de esos elementos fueron filtradas mediante una bomba neumática ANTLIA, modelo SP 050/2, con tres portafiltros de idéntica marca y modelo FP050/0, todo ello construido en policarbonato. Para esta operación se utilizaron filtros de 2, 0.45 y 0.1 μm sucesivamente dispuestos. La muestra filtrada se almacenó en botes de polietileno de 100 ml, previamente lavados, homogeneizados en soluciones de ácido nítrico ultrapuro durante varios días, enjuagados posteriormente con agua destilada y secados en corriente de aire. Las alícuotas tomadas se acidificaban con ácido nítrico ultrapuro hasta pH < 1 y la totalidad del equipo era cuidadosamente lavada con agua destilada antes de acceder al siguiente punto de desmuestre.

Tabla 1: Contenidos de Tierras Raras en muestras seleccionadas de aguas ácidas a lo largo del curso del Arroyo del Val (Provincia de Zaragoza). Concentraciones y límites de detección del ICP-MS expresados en ppm.

	Límites de detección	W-16	W-9	W-5
La	$1\cdot10^{-4}$	0.0119	0.0074	0.0049
Ce	$1\cdot10^{-4}$	0.0345	0.0209	0.0133
Pr	$1\cdot10^{-4}$	0.006	0.0034	0.0022
Nd	$1\cdot10^{-4}$	0.0359	0.0188	0.0117
Sm	$1\cdot10^{-4}$	0.0117	0.0067	0.0040
Eu	5.10^{-5}	0.0034	0.00186	0.00118
Gd	$1\cdot10^{-4}$	0.0174	0.0083	0.0053
Tb	$1\cdot10^{-4}$	0.0022	0.00120	0.0004
Dy	$1\cdot10^{-4}$	0.0101	0.0058	0.0038
Ho	$5\cdot10^{-5}$	0.00167	0.0009	0.00054
Er	$1\cdot10^{-4}$	0.0039	0.002	0.0013
Tm	$1\cdot10^{-4}$	0.0004	0.0002	0.0001
Yb	$1\cdot10^{-4}$	0.0024	0.0013	0.0007
Lu	5.10^{-5}	0.00034	0.0001	0.00009
Conduct. (μS/cm)		839.0	721.0	625.0
pH		3.32	4.39	4.64

Las Tierras Raras fueron analizadas mediante ICP-MS, con los límites de detección y resultados señalados en la tabla 1. En esta tanda de análisis no se utilizaron técnicas de preconcentración para la determinación de estos elementos, ya que las elevadas concentraciones en las que se encuentran en las soluciones ácidas justifican sobradamente esta simplificación de la metodología.

Al valorar los resultados hay que tener presente que tanto la concentración como las pautas de REE en solución están condicionadas directamente por la metodología de filtración seguida en el muestreo. El efecto de las técnicas de filtración en la separación de la fracción coloidal, en suspensión y en solución verdadera es ambiguo y problemático (Sholkovitz, 1992). En la bibliografía sobre el tema no existe un criterio fijo para hablar de fracciones particuladas, disueltas, coloidales o en solución verdadera y, mucho menos, una sistematización de los tamaños de filtro que permiten separarlas. En cualquier caso, el tamaño mínimo de poro utilizado en la metodología de filtración de este estudio (0.1 μm) permite considerar que los resultados obtenidos corresponden, cuando menos, a la fracción disuelta (considerando como tal la constituida por la fracción en solución verdadera y la fracción coloidal; Sholkovitz, 1992)

Los contenidos en REE y sus modelos de distribución en el presente trabajo han sido normalizados frente a NASC (North American Shale Composite; Haskin *et al.*, 1968; Gromet *et al.*, 1984).

3. Caracteres generales del sistema y evolución de los contenidos en Tierras Raras

Las pautas de evolución geoquímica de las aguas ácidas del Arroyo del Val se encuentran condicionadas por los sucesivos aportes de los afluentes con soluciones neutras (valores de pH entre 7 y 8) y algo más diluidas (con conductividades entre 500 y 600 μS/cm; p. ej. Auqué *et al.*, 1994) y por los subsiguientes fenómenos de mezcla y floculación coloidal. Estos procesos determinan una progresiva disminución de las concentraciones de casi todos los componentes en solución a lo largo del curso del arroyo; disminución que se evidencia, especialmente, en los elementos metálicos, retenidos en los fenómenos de floculación por coloides constituidos fundamentalmente por oxihidróxidos, sulfatos básicos de aluminio y hierro, y silicatos hidratados de aluminio, amorfos (Gimeno, 1991; Auqué *et al.*, 1993).

Los rasgos más característicos de esta evolución pueden quedar representados por la variación del pH y conductividad de las soluciones ácidas a lo largo del tramo muestreado. En la figura 2 se ha representado los valores de estos parámetros en la dirección de flujo constatándose, como consecuencia de los procesos anteriormente mencionados, un progresivo incremento del pH y disminución de la conductividad de las aguas. Tendencias similares han sido definidas para este sistema en muestreos más amplios realizados sobre la totalidad del curso del Arroyo (Gimeno 1991; Gimeno *et al.*, 1994).

En este contexto geoquímico, las concentraciones de REE determinadas en las soluciones ácidas del arroyo, independientemente de su situación a lo largo del curso, son considerablemente elevadas (tabla 1) y muy superiores a las determinadas en aguas marinas o en soluciones continentales de carácter neutro-básico. Las únicas soluciones con contenidos parangonables o superiores a las del Val en la bibliografía corresponden a sistemas hídricos de distinto tipo, pero todos ellos caracterizados por soluciones de pH marcadamente ácido (p. ej. Fee *et al.*, 1992; Miekeley *et al.*, 1992).

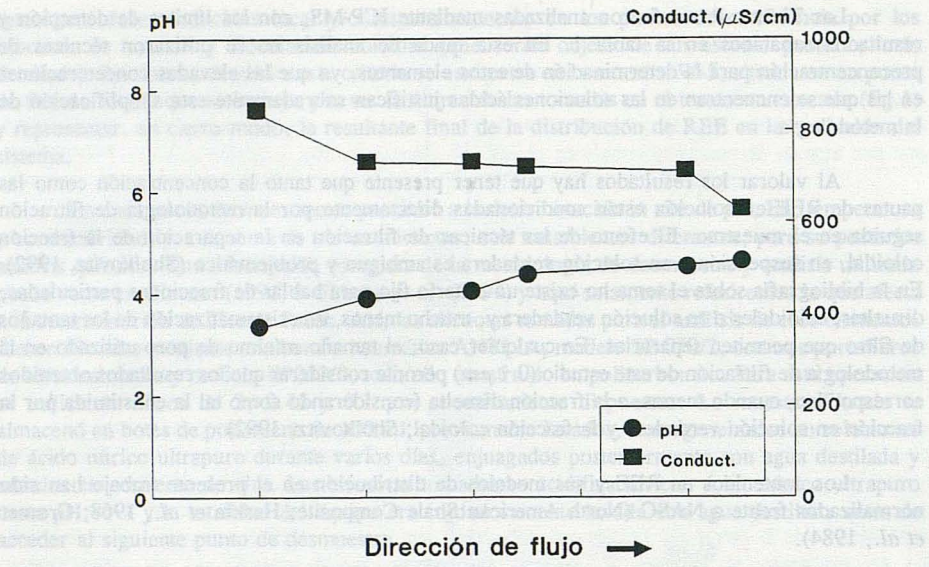


Figura 2. Evolución de los valores de pH y conductividad de las aguas ácidas del Arroyo del Val en la dirección de flujo.

Teniendo en cuenta la disposición espacial de las muestras, puede apreciarse una marcada disminución en los contenidos de toda la serie de las Tierras Raras (tabla 1) en la dirección de flujo: las concentraciones de la muestra W-16 (correspondiente a la parte más alta muestreada en el arroyo) triplican las de la W-5 (última muestra tomada aguas abajo; figura 1), situándose las concentraciones del resto de muestras entre esos dos extremos (figura 3 B).

Los contenidos en REE analizados tras cada punto de mezcla y floculación coloidal son sistemáticamente menores que los existentes antes de producirse ese proceso. El único tramo en el que las concentraciones de Tierras Raras se mantienen prácticamente constantes es el existente entre las muestras W-10 y W-9 (figura 3 A y B), en el que precisamente no se produce ningún tipo de fenómeno de mezcla de aguas ni se aprecian fenómenos de floculación (figura 1). Estos procesos de floculación coloidal son los responsables, por tanto, de la disminución de concentración de las Tierras Raras a lo largo del curso del Arroyo del Val representando, como sucedía con los elementos metálicos (Gimeno *et al.*, 1994), un efectivo mecanismo de retención para esos elementos.

4. Evolución en la pauta de distribución de Tierras Raras

Independientemente de las diferencias de concentración existentes entre las muestras analizadas, los modelos de distribución de REE, normalizados frente a NASC, presentan una característica común (figura 3 B): todos ellos muestran una marcada convexidad en torno a las Tierras Raras intermedias (IREE; Eu, Gd y Tb), resultado de un importante enriquecimiento de este grupo de las Tierras Raras respecto a las ligeras (LREE) y pesadas (HREE).

Pautas similares a las descritas han sido citadas en soluciones ácidas de distintos sistemas hídricos de baja temperatura (p. ej. Elderfield *et al.*, 1990; Smedley, 1991; Fee *et al.*, 1992; Gosselin *et al.*, 1992) y atribuidas tanto a la fracción coloidal transportada por las soluciones

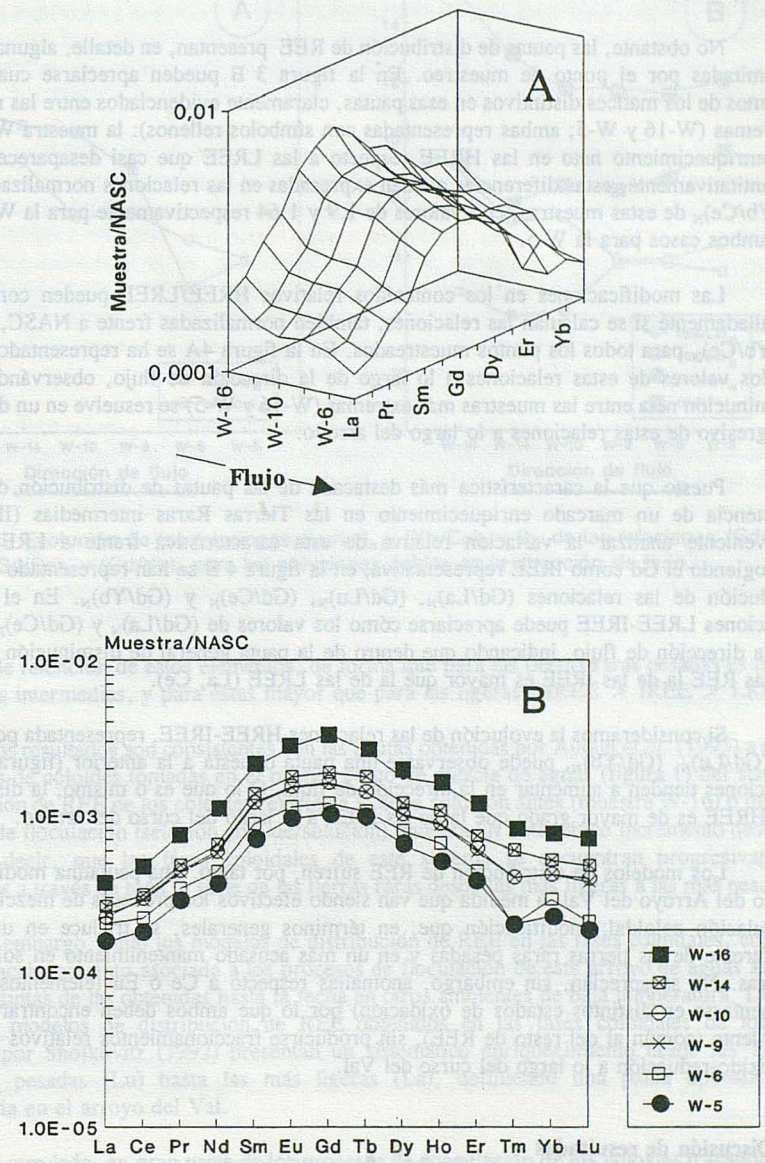


Figura 3. A.- Representación (en escala logarítmica) de la evolución de los modelos de distribución de REE (normalizados frente a NASC) en la dirección de flujo. B.- Representación de los modelos de distribución para todas las muestras analizadas, en una sección perpendicular a la dirección de flujo.

(Elderfield *et al.*, 1990) como a la interacción de las soluciones con materiales típicamente enriquecidos en IREE (arcillas y oxihidróxidos de hierro; Gosselin *et al.*, 1992; Johannesson y Lyons, 1993).

No obstante, las pautas de distribución de REE presentan, en detalle, algunas diferencias delimitadas por el punto de muestreo. En la figura 3 B pueden apreciarse cualitativamente algunos de los matices distintivos en esas pautas, claramente evidenciados entre las muestras más extremas (W-16 y W-5; ambas representadas con símbolos rellenos): la muestra W-16 presenta un enriquecimiento neto en las HREE respecto a las LREE que casi desaparece en la W-5. Cuantitativamente, estas diferencias quedan expresadas en las relaciones normalizadas $(Lu/La)_N$ e $(Yb/Ce)_N$ de estas muestras, con valores de 1.9 y 1.64 respectivamente para la W-16 y de 1.2 en ambos casos para la W-5.

Las modificaciones en los contenidos relativos HREE/LREE pueden constatarse más detalladamente si se calculan las relaciones, también normalizadas frente a NASC, de $(Lu/La)_N$ y $(Yb/Ce)_N$ para todos los puntos muestreados. En la figura 4A se ha representado la variación en los valores de estas relaciones a lo largo de la dirección de flujo, observándose cómo la disminución neta entre las muestras más extremas (W-16 y W-5) se resuelve en un decrecimiento progresivo de estas relaciones a lo largo del arroyo.

Puesto que la característica más destacada de las pautas de distribución de REE es la existencia de un marcado enriquecimiento en las Tierras Raras intermedias (IREE) resulta conveniente analizar la variación relativa de esta característica frente a LREE y HREE. Escogiendo el Gd como IREE representativa, en la figura 4 B se han representado las pautas de evolución de las relaciones $(Gd/La)_N$, $(Gd/Lu)_N$, $(Gd/Ce)_N$ y $(Gd/Yb)_N$. En el caso de las relaciones LREE-IREE puede apreciarse cómo los valores de $(Gd/La)_N$ y $(Gd/Ce)_N$ disminuyen en la dirección de flujo, indicando que dentro de la pauta general de disminución del conjunto de las REE la de las IREE es mayor que la de las LREE (La, Ce).

Si consideramos la evolución de las relaciones HREE-IREE, representada por los valores de $(Gd/Lu)_N$, $(Gd/Yb)_N$, puede observarse una pauta opuesta a la anterior (figura 4B): ambas relaciones tienden a aumentar en la dirección de flujo. O lo que es o mismo, la disminución de las HREE es de mayor grado que la de las IREE a lo largo del curso del arroyo.

Los modelos de distribución de REE sufren, por tanto, una paulatina modificación a lo largo del Arroyo del Val, a medida que van siendo efectivos los procesos de mezcla de aguas y floculación coloidal; modificación que, en términos generales, se traduce en una remoción preferente de las tierras raras pesadas y en un más acusado mantenimiento en solución de las ligeras. No se aprecian, sin embargo, anomalías respecto a Ce o Eu (elementos que pueden presentarse en distintos estados de oxidación) por lo que ambos deben encontrarse en estado trivalente (común al del resto de REE), sin producirse fraccionamientos relativos por procesos de óxido-reducción a lo largo del curso del Val.

5. Discusión de resultados

La evolución descrita en los modelos de distribución de REE a lo largo del Arroyo del Val evidencia un apreciable fraccionamiento asociado a los fenómenos de floculación coloidal efectivos en el sistema. Este fraccionamiento se encuentra caracterizado por una distinta intensidad en los procesos de remoción de REE de las soluciones, progresivamente mayor desde las tierras raras ligeras hasta las pesadas. Ello implica que los coloides presentan una distinta

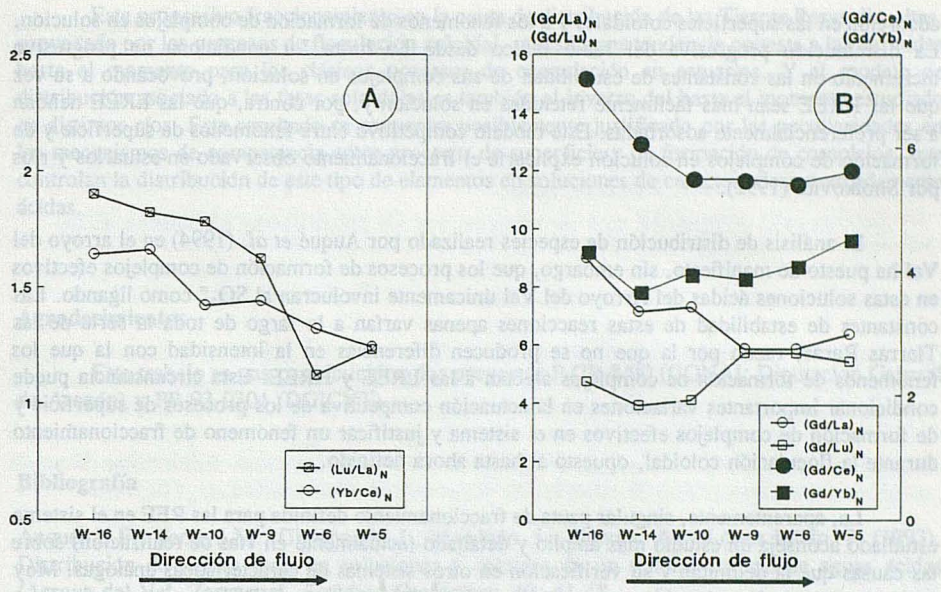


Figura 4. A.- Evolución de las relaciones $(Lu/La)_N$ y $(Yb/Ce)_N$ y B.- de las relaciones $(Gd/La)_N$, $(Gd/Lu)_N$, $(Gd/Ce)_N$ y $(Gd/Yb)_N$ para las soluciones ácidas en la dirección de flujo.

capacidad de retención de estos elementos, de forma que para las tierras raras pesadas es mayor que para las intermedias, y para éstas mayor que para las ligeras (HREE > IREE > LREE).

Estos resultados son consistentes con las pautas obtenidas por Auqué *et al.* (1993) a partir de muestras de coloides tomadas en el primer punto de mezcla de aguas (figura 1) del sistema: la distribución de REE de los coloides referida a la de la solución antes (muestra W-16) o durante el proceso de floculación (relación coloide/solución) muestra un sistemático incremento desde La a Lu. Es decir, que las fases coloidales de este sistema se encuentran progresivamente enriquecidas a través de toda la serie de las tierras raras desde las más ligeras a las más pesadas.

Sin embargo, tanto los modelos de distribución de REE en las fases coloidales, como la pauta de fraccionamiento asociada a los procesos de floculación de este arroyo de aguas ácidas, resultan distintas de las obtenidas hasta la fecha en otros ambientes de baja temperatura. De esta forma, los modelos de distribución de REE obtenidos en las fases coloidales de los ríos estudiados por Sholkovitz (1992) presentan un sistemático enriquecimiento desde las Tierras Raras más pesadas (Lu) hasta las más ligeras (La), definiendo una pauta opuesta a la caracterizada en el arroyo del Val.

Por otro lado, en gran parte de los procesos de coagulación de los coloides transportados por los ríos en los estuarios (inducidos por la mezcla con agua marina al llegar a la desembocadura) se produce una retención preferente de las LREE frente a las HREE (Sholkovitz, 1992); carácter que resulta coherente con la existencia de coloides enriquecidos en LREE pero, por ello, también opuesta al determinado en el arroyo del Val.

Los estudios realizados hasta el momento indican que los procesos de fraccionamiento de las REE en solución se deben a sus diferentes afinidades relativas frente a los procesos de

adsorción en las superficies coloidales y a los fenómenos de formación de complejos en solución. La disminución progresiva del radio iónico desde La hasta Lu condiciona un progresivo incremento en las constantes de estabilidad de sus complejos en solución, provocando a su vez que las HREE sean más fácilmente retenidas en solución y, por contra, que las LREE tiendan a ser preferencialmente adsorbidas. Este modelo competitivo entre fenómenos de superficie y de formación de complejos en solución explicaría el fraccionamiento observado en estuarios y ríos por Sholkovitz (1992).

El análisis de distribución de especies realizado por Auqué *et al.* (1994) en el arroyo del Val ha puesto de manifiesto, sin embargo, que los procesos de formación de complejos efectivos en estas soluciones ácidas del Arroyo del Val únicamente involucran al $SO_4^=$ como ligando. Las constantes de estabilidad de estas reacciones apenas varían a lo largo de toda la serie de las Tierras Raras, razón por la que no se producen diferencias en la intensidad con la que los fenómenos de formación de complejos afectan a las LREE y HREE. Esta circunstancia puede condicionar importantes variaciones en la actuación competitiva de los procesos de superficie y de formación de complejos efectivos en el sistema y justificar un fenómeno de fraccionamiento durante la floculación coloidal, opuesto al hasta ahora definido.

La, aparentemente, singular pauta de fraccionamiento definida para las REE en el sistema estudiado aconseja un estudio más amplio y detallado (actualmente en vías de realización) sobre las causas que la delimitan y su verificación en otros sistemas de características análogas. Muy recientemente, se han obtenido pautas similares de fraccionamiento de REE en sistemas de aguas ácidas análogos al del Arroyo del Val (D.K. Nordstrom, U.S. Geol. Surv.; com. pers.) lo que evidencia la importancia del estudio de este tipo de sistemas para una más amplia comprensión de los procesos que controlan la distribución de REE en los sistemas naturales.

6. Conclusiones

Los contenidos de REE transportados por las soluciones ácidas del Arroyo del Val son considerablemente elevados, apreciándose una marcada disminución en la dirección de flujo como consecuencia de los sucesivos procesos de floculación coloidal. Sin embargo, esta disminución en las concentraciones no afecta homogéneamente a todo el conjunto de las Tierras Raras; si bien las pautas de distribución de REE presentan en todos los casos un característico enriquecimiento en las Tierras Raras intermedias, resultan apreciables modificaciones progresivamente efectivas en los contenidos relativos de HREE, IREE y LREE a lo largo del curso del arroyo.

Inicialmente el modelo de distribución de REE está caracterizado por un empobrecimiento en las tierras raras más ligeras desde Gd a La y en las más pesadas desde Gd a Lu, más marcado en el primer caso que en el segundo. Conforme se suceden los procesos de floculación las diferencias en el grado de empobrecimiento de HREE y LREE respecto al Gd tienden a igualarse, existiendo por tanto una retención preferente de las tierras raras pesadas frente a las ligeras en el proceso.

Más en detalle, la evolución de la distribución de REE a lo largo del curso del arroyo evidencia la presencia de una gradación en la remoción generalizada de estos elementos de forma que HREE > IREE > LREE. Circunstancia que resulta consistente con el modelo de distribución de REE obtenidos en muestras de coloides (normalizados respecto a los de la solución con la que están en contacto) y caracterizados por un paulatino enriquecimiento desde La hasta Lu.

Este progresivo fraccionamiento en la pauta de distribución de las Tierras Raras disueltas, provocado por los procesos de floculación coloidal, muestra una tendencia opuesta a las definidas hasta el momento para los clásicos procesos de coagulación en estuarios. Y el modelo de distribución asociado a las fases coloidales es también el inverso del hasta el momento estudiado en distintos ríos. Este resultado se encuentra posiblemente justificado por las peculiaridades en los mecanismos de competencia entre procesos de superficie y de formación de complejos, que controlan la distribución de este tipo de elementos en soluciones de características marcadamente ácidas.

Agradecimientos

Este trabajo es una contribución a los proyectos P CB-8/90 (CONAI; Diputación General de Aragón) y PB 93-0304 (DGICYT).

Bibliografía

Auqué, L.F.; Tena, J.M.; Gimeno, M.J.; Mandado, J.; Zamora, A. y López Julián, P. (1993). Distribución de tierras raras en soluciones y coloides de un sistema natural de aguas ácidas (Arroyo del Val, Zaragoza). *Estudios Geológicos*, 49, 41-48.

Auqué, L.F.; Tena, J.M.; Gimeno, M.J.; Mandado, J.; López Julián, P. y Zamora, A. (1994). Especiación de Tierras Raras en las soluciones ácidas y neutras del sistema de drenaje del Arroyo del Val (Zaragoza). *Estudios Geológicos* (en prensa).

De Baar, H.J.W.; German, C.R.; Elderfield, H. & Van Gaans (1988). Rare earth element distributions in anoxic waters of the Cariaco Trench. *Geochim. Cosmochim. Acta*, 52, 1203-1220.

Elderfield, H; Upstill-Goddard, R. & Sholkovitz, E.R. (1990). The rare earth elements in rivers, estuaries and coastal seas and their significance to the composition of ocean waters. *Geochim. Cosmochim. Acta*, 54, 971-991.

Fee, J.A.; Gaudette, H.E.; Berry Lyons, W. & Long, D.T. (1992). Rare-earth element distribution in Lake Tyrrell groundwaters, Victoria, Australia. *Chem. Geol.*, 96, 67-93.

Gimeno, M. J. (1991). *Ambiente geológico y caracterización del sistema geoquímico "Arroyo del Val": relaciones entre fases minerales y especies en disolución.* Tesis de Licenciatura, Universidad de Zaragoza, 184 pp. (No publicada).

Gimeno, M.J.; Tena, J.M.; Auqué, L.F. y Mandado, J. (1994). Caracterización geoquímica del sistema de aguas ácidas del Arroyo del Val (Zaragoza). *Bol. R. Soc. Esp. Hist. Nat. (Sección Geología)*, 89, 5-17.

Goldstein, S.J. & Jacobsen, S.B. (1988). Rare earth elements in river waters. *Earth Planet. Sci. Lett.*, 89, 35-47.

Gosselin, D.C.; Smith, M.R.; Lepel, E.A. & Laul, J.C. (1992). Rare earth elements in chloride-rich groundwater, Palo Duro Basin, Texas, USA. *Geochim. Cosmochim. Acta*, 56, 1495-1505.

Gromet, L.P.; Dymek, R.F.; Haskin, L.A. & Korotev, R.L. (1984). The "North American Shale Composite": its compilation, major and trace element characteristics. *Geochim. Cosmochim. Acta*, 48, 2469-2482.

Haskin, L.A.; Haskin, M.A.; Frey, F.A. & Wildeman, T.R. (1968). Relative and absolute terrestrial abundances of the rare earths. In: *Origin and Distribution of the Elements* (L.H. Ahrens, ed.), pp. 889-911. Pergamon Press.

Johannesson, K.H. & Lyons, W.B. (1993). Rare earth element concentrations and speciation in acidic and alkaline natural waters. *Geological Society of America. 1993 Annual Meeting, Boston, Massachusetts*. October, 25-28, 1993. A-254.

Miekeley, N.; Couthino de Jesus, H.; Porto da Silveira, C.L.; Linsalata, P. and Morse, R. (1992). Rare-earth elements in groundwaters from the Osamu Utsumi mine and Morro do Ferro analogue study sites, Poços de Caldas, Brazil. *Jour. Geochem. Explor.*, 45, 365-387.

Sholkovitz, E.R. (1992). Chemical evolution of rare earth elements: fractionation between colloidal and solution phases of filtered river water. *Earth Planet. Sci. Let.*, 114, 77-84.

Sholkovitz, E.R. (1993). The geochemistry of rare earth elements in the Amazon River estuary. *Geochim. Cosmochim. Acta,* 57, 2181-2190.

Smedley, P.L. (1991). The geochemistry of rare earth elements in groundwater from the Carnmenellis area, southwest England. *Geochim. Cosmochim. Acta,* 55, 2767-2779.

Rev. Academia de Ciencias. Zaragoza. 49 (1994)

Importancia de la variación de solubilidad de la mirabilita con la temperatura en la evolución geoquímica de las lagunas de Los Monegros (Zaragoza).

L.F. Auqué[1], V. Vallès[2], H. Zouggari[3], P.L. López[1] y G. Bourrié[3]

(1) Área Petrología y Geoquímica. Depto. Ciencias de la Tierra. Fac. Ciencias. Universidad de Zaragoza. 50009 ZARAGOZA (España).
(2) Laboratoire de Science du Sol, Institut National de la Recherche Agronomique. Domaine St Paul, B.P. 91, 84143 Montfavet CEDEX (France).
(3) Laboratoire de Science du Sol, Institut National de la Recherche Agronomique. 65 Route de St Brieuc, 35 000 Rennes (France).

Abstract

The geochemical evolution of Los Monegros playa-lakes is affected by temperature fluctuations in the brine body at different time scales. Temperature shifts promote seasonally (monthly), daily and even during minor cycles mineralogical and compositional changes, mainly related with mirabilite crystallization. Experimental determination on mirabilite solubility between 0 and 30 °C, using natural brines, allow us to see the importance of this non-isothermal aspect of brine evolution.

Seasonal mirabilite precipitation cycles occur in Cl-SO_4-Na-(Mg) brines of Los Monegros: precipitated mirabilite during winter dissolves during summer. But daily cycles also exist during winter-spring brine evolution: temperature oscillations of 10 °C in the 0-30 °C interval (frequent in the studied area during spring) promote important changes in mirabilite solubility. So, mirabilite crystallization occurs during spring nights by lowering temperature, whereas diurnal temperatures promote their dissolution.

When high saturation levels are reached by evaporative concentration and the amount of precipitated mirabilite is important in the system, diurnal temperature fluctuations induce quick mirabilite-solution reequilibrium processes. At this moment, temperature changes between 20 and 30 °C are common in the lakes and modifications of mirabilite solubility produce their maximun effects on compositional characters of solutions in this temperature interval: several hundreds of grams/kg water of mirabilite are movilized, brine ionic strength changes from 2 to 8 molal and water activity varies from 0.966 to 0.896. The gypsum equilibrium state reached by evaporative concentration of brines holds without additional mass transfer processes in spite of the important compositional changes induced by mirabilite reequilibrium and the common ion effect.

1. Introducción.

La mayoría de los lagos salinos en zonas endorreicas continentales presentan marcados gradientes de temperatura en la vertical y/o se ven sometidos a apreciables variaciones diarias, mensuales y/o estacionales de este parámetro, con importantes efectos en los procesos de

cristalización de fases minerales y en el quimismo de las salmueras. No obstante, este aspecto no isotermo en la evolución de este tipo de sistemas ha sido frecuentemente obviado y los estudios que lo han considerado (p. ej. Pueyo, 1978-79; Smith *et al.*, 1987) han evidenciado su incidencia a partir de observaciones mineralógicas directas, sin establecer con precisión su relación con las variables fisicoquímicas y composicionales de la salmuera.

El estudio de estas relaciones requiere la utilización de modelos termodinámicos adecuados para el tratamiento de soluciones altamente concentradas dentro de metodologías de modelización geoquímica. Y hasta hace poco, el modelo más frecuentemente utilizado era el tabulado a 25 °C por Harvie, Moller y Weare (modelo HMW; Harvie & Weare, 1980; Harvie *et al.*, 1984) a partir de la aproximación de interacciones iónicas de Pitzer (1973, 1979). Sin embargo, las recientes ampliaciones propuestas por Plummer *et al.* (1988), Moller (1988) o Greenberg & Moller (1989) para este modelo a temperaturas distintas de 25 °C permiten abordar el tratamiento de procesos a temperatura variable en la evolución geoquímica de salmueras.

Las lagunas saladas de Los Monegros - Bajo Aragón constituyen uno de los sistemas endorreicos en los que se ha evidenciado los efectos de las variaciones estacionales de temperatura sobre el proceso de cristalización de mirabilita a partir de salmueras libres (Pueyo, 1978-79; Pueyo e Inglés, 1987). Las observaciones de campo han puesto de manifiesto, además, que los efectos de las variaciones de temperatura sobre este proceso reducen su escala de influencia a intervalos mensuales e incluso diarios cuando la salmuera alcanza un elevado estado de concentración por evaporación.

Este sistema endorreico resulta, por tanto, especialmente apropiado para analizar cuantitativamente las modificaciones composicionales de la salmuera inducidas por procesos heterogéneos especialmente sensibles a la temperatura (caso de la precipitación/disolución de mirabilita). Lógicamente estas variaciones composicionales se encuentran superpuestas a la pauta general de evolución de la salmuera por concentración evaporativa. Por ello se ha preparado un dispositivo experimental en condiciones controladas de laboratorio, utilizando muestras minerales y de salmuera tomadas del sistema natural, que permita valorar y aislar los efectos producidos por la variación de temperatura de los inducidos por la progresiva concentración de las lagunas.

En este trabajo se analiza la influencia de la temperatura sobre los procesos de precipitación/disolución de mirabilita en la salmuera, las modificaciones composicionales directamente relacionadas con la variación de solubilidad de este mineral así como los efectos inducidos sobre la estabilidad del yeso, una de las fases minerales más abundantes en este sistema. Todo ello a través de la utilización de una de las ampliaciones propuestas del modelo HMW para el tratamiento de sistemas altamente concentrados y en condiciones no isotermas.

2. Localización del área de estudio. Características generales.

El conjunto de lagunas saladas distribuidas en la zona central y meridional de la Cuenca del Ebro (comarcas de Los Monegros y Bajo Aragón) constituye, junto con el de La Mancha, uno de los sistemas endorreicos más importantes de la Península (Pueyo y De La Peña, 1991). La elevada aridez de este área (con una pluviometría inferior a 300 mm/año) junto con la existencia de vientos dominantes secos ("cierzo", de componente NW) y una elevada tasa de insolación favorecen la evaporación de las lagunas.

Las salmueras de estas lagunas son mayoritariamente de tipo Cl-SO_4-Na-(Mg) (Pueyo, 1978-79; Mingarro *et al.*, 1981; Pueyo e Inglés, 1987) y la mineralogía ligada a los procesos de

precipitación por evaporación de las aguas presenta un carácter estacional: durante la primavera-verano se produce una secuencia de precipitación constituida por carbonatos, yeso y halita y durante el invierno se produce la precipitación de mirabilita (conocida en la zona como "sal de invierno"; Pueyo, 1978-79).

A parte del referido ciclo estacional en la génesis de mirabilita, distintas observaciones evidencian la existencia de ciclos de menor escala (mensual o diaria). Así, Pueyo e Inglés (1987) señalan que las oscilaciones diarias de temperatura inducen la precipitación de mirabilita en las noches primaverales; observación que resulta justificable considerando que las variaciones diarias de temperatura en esa estación alcanzan frecuentemente los 20 °C (en los últimos años, los promedios de temperatura máxima y mínima durante los meses de Abril y Mayo en la zona alcanzan valores entre los 5 y 30 °C).

Por otro lado, en las campañas de muestreo realizadas en la zona (Vallès *et al.*, 1994) durante la primavera, cuando la salmuera alcanza niveles de concentración evaporativa suficientes como para generar importantes depósitos de mirabilita, se observan efectos todavía más drásticos inducidos por las variaciones de temperatura. En muestras de la salmuera aisladas de su entorno en botes de polietileno se observaban espectaculares procesos de precipitación de este mineral cuando simplemente eran colocados a la sombra y su temperatura descendía unos cuantos grados centígrados. Colocadas de nuevo al sol, los cristales de mirabilita volvían a disolverse aunque, aparentemente, con una cinética más lenta que la de precipitación. Es decir que, llegados a este punto de la secuencia evaporativa de la salmuera, los fenómenos de reequilibrio de la solución respecto a la mirabilita se hacen especialmente sensibles a las variaciones de temperatura y operan con una cinética considerablemente rápida.

Uno de los puntos en los que se han verificado ese conjunto de observaciones es la laguna del Salicar, situada al sur de Bujaraloz (Zaragoza). La mineralogía de los depósitos de esta laguna está representada fundamentalmente por yeso y, eventualmente (en la etapa de invierno-primavera), por mirabilita. El esquema general de evolución geoquímica identificado para esta laguna responde a una secuencia de precipitación de carbonatos, yeso y mirabilita (Vallès *et al.*, 1994) dentro de la pauta de evaporación progresiva de la salmuera.

En esta laguna se procedió al muestreo de la salmuera, en primavera, cuando la solución ya presentaba unos niveles de concentración elevados (saturación respecto a yeso y mirabilita), así como al desmuestre de la abundante mirabilita generada en la laguna a partir del proceso evaporativo de la salmuera. Los análisis realizados mediante DRX sobre las muestras sólidas señalan la presencia casi exclusiva de mirabilita sin que aparezcan indicaciones apreciables de otras fases minerales.

3. Metodología.

Las experiencias han sido realizadas a partir de las muestras de mirabilita y salmuera extraídas del sistema natural. Para ello se introdujo el material (mirabilita + salmuera) en un recipiente de polietileno de 500 mls y se colocó en un baño maría capaz de mantener unas condiciones isotermas con una precisión de 0.1 °C, analizándose la variación de la composición de la salmuera en un rango de temperatura creciente de 0 a 30 °C, a intervalos de 5 °C. Para cada uno de estos intervalos la temperatura se mantenía constante durante 24 horas (salvo en el caso del primero, a 0 °C, que se mantuvo durante 48 horas) antes de proceder a la toma de muestra para su análisis.

Una vez alcanzada la temperatura de 30 °C se invirtió el sentido del experimento, disminuyendo la temperatura hasta 0 °C también en intervalos de 5 °C, a fin de detectar eventuales efectos cinéticos en las reacciones de disolución/precipitación involucradas. La toma de muestras en los intervalos de temperatura investigados fue realizada con una microbureta HAMILTON, programada para extraer 250 microlitros de solución. La densidad fue determinada por pesada analizándose posteriormente Na^+, K^+, Ca^{2+}, Mg^{2+}, $SO_4^=$ y Cl^- mediante cromatografía iónica, tras efectuar la dilución de las muestras.

La calidad de los resultados analíticos obtenidos en el experimento ha sido verificada mediante el cálculo del balance de cargas, según la expresión de error de balance (E.B.):

$$E.B. = \frac{\sum_{cationes} - \sum_{aniones}}{(\sum_{cationes} + \sum_{aniones}) / 2}$$

Los resultados analíticos presentados en la tabla 1, tanto para la experiencia a temperatura creciente como decreciente, muestran errores de balance inferiores a 0.1 salvo en el caso de las muestras tomadas a temperaturas de 10 y 20 °C, en las que se alcanzan valores de 0.11 a 0.15.

Dado el escaso volumen de muestra manejado no ha sido posible determinar la alcalinidad. No obstante, la alcalinidad carbonatada determinada en las aguas de la Laguna del Salicar (Rezagui, 1993; Vallès et al., 1994) presenta valores considerablemente bajos (entre $9 \cdot 10^{-4}$ y $3.3 \cdot 10^{-2}$ eq/kg) frente a las concentraciones iónicas del resto de elementos. Introduciendo en los cálculos de balance el valor máximo de alcalinidad medido en esos muestreos de la laguna se obtienen resultados que no varían más allá de 0.02 unidades sobre los inicialmente obtenidos. Lo que señala que la incertidumbre introducida por la ausencia de datos de alcalinidad en el cálculo de error de balance es mínima.

Las elevadas concentraciones de las soluciones analizadas impiden la utilización de los clásicos modelos de Asociación Iónica basados en la ecuación de Debye-Hückel para determinar las actividades de las especies disueltas (Garcés *et al.*, 1991, 1992) y calcular con precisión los productos de actividad iónica correspondientes a las fases minerales involucradas.

Por ello, para efectuar estos cálculos se ha utilizado el código PHRQPITZ (Plummer *et al.*, 1988) que incorpora y amplía a rangos de temperaturas entre 0 y 60 °C el modelo de interacciones iónicas o de coeficientes viriales de Pitzer (1973, 1979) con los parámetros obtenidos por Harvie y colaboradores (Harvie & Weare, 1980; Harvie *et al.*, 1984) para el sistema $Na\text{-}K\text{-}Mg\text{-}Ca\text{-}H\text{-}Cl\text{-}SO_4\text{-}OH\text{-}HCO_3\text{-}CO_3\text{-}CO_2\text{-}H_2O$, a 25 °C. El código utiliza las concentraciones de los componentes analizados para calcular las actividades iónicas individuales y determinar el grado de saturación de la solución respecto a distintas fases minerales. Si bien el modelo original propuesto por Harvie y colaboradores, a 25 °C, ha sido frecuentemente utilizado en el tratamiento de sistemas naturales altamente concentrados, no ocurre lo mismo con la propuesta de ampliación a distintas temperaturas incluida en el PHRQPITZ y sólo recientemente se han comenzado a explorar sus posibilidades (p. ej. Bischoff *et al.*, 1991, 1993).

Los resultados de este tratamiento termodinámico y la verificación de las situaciones de equilibrio se realizarán comparando los productos de actividad iónica (P.A.I.) calculados respecto a las constantes de equilibrio o productos de solubilidad de las fases minerales consideradas (K(T)) o bien, de forma más sintética, a través del cálculo de índice de saturación expresado según la ecuación:

Tabla 1.- Resultados de las experiencias de solubilidad de mirabilita a temperatura creciente y decreciente. Las concentraciones están expresadas en moles/kg de agua.

Experiencia a temperatura creciente						
TEMP.	$SO_4^=$	Cl^-	Na^+	K^+	Ca^{2+}	Mg^{2+}
0	0.47044	1.09172	1.41583	0.04980	0.02002	0.35987
5	0.48501	0.95569	1.36677	0.03001	0.01722	0.31106
10	0.81088	0.91577	1.59888	0.02889	0.01630	0.29501
15	0.81873	0.93727	1.95716	0.02714	0.01308	0.27980
20	1.05908	0.79924	2.48368	0.02413	0.01308	0.00968
25	1.64076	0.71568	3.57167	0.02475	0.01347	0.22078
30	2.48397	0.62197	4.87097	0.02187	0.01437	0.16318
Experiencia a temperatura decreciente						
25	1.73649	0.74905	3.63538	0.02454	0.01618	0.23001
20	1.09506	1.29230	2.40898	0.02658	0.02205	0.25106
15	0.89021	0.97212	2.22131	0.03228	0.02179	0.32524
10	0.51368	0.92259	1.50175	0.03064	0.02035	0.31095
5	0.39722	0.89207	1.19839	0.03064	0.01525	0.30350
0	0.23920	0.84337	0.81000	0.02640	0.01818	0.27165

$$I.S. = \log \frac{P.A.I.}{K(T)}$$

de forma que valores de I.S. mayores que 0 indican que la solución se encuentra sobresaturada respecto a la fase mineral en cuestión; valores inferiores a 0 corresponden a situaciones de subsaturación; y valores de I.S. = 0 corresponden a situaciones de equilibrio termodinámico mineral-solución.

La precisión de los resultados de I.S. (o lo que es lo mismo, los rangos de incertidumbre producidos por la propagación de imprecisiones termodinámicas y analíticas) obtenidos con modelos de interacción iónica no se encuentran todavía correctamente establecidos. En los escasos trabajos en los que se hace referencia al tema se utiliza la banda de error definida por Jenne *et al.* (1980), en torno al 5% del log K para la fase mineral considerada (p. ej. Langmuir & Melchior, 1985; Fisher & Kreitler, 1987). Se trata de una aproximación algo arbitraria y no totalmente rigurosa que llevaría a considerar un rango razonable de $\pm$ 0.2 unidades en el caso de yeso y anhidrita pero excesivamente estrecho en el de la mirabilita ($\pm$ 0.06 unidades de I.S.). Una revisión de los valores de la constante de equilibrio propuestas en la bibliografía para la mirabilita evidencia fácilmente la existencia de diferencias de $\pm$ 0.15 unidades, más del doble de la obtenida a partir de aquella aproximación.

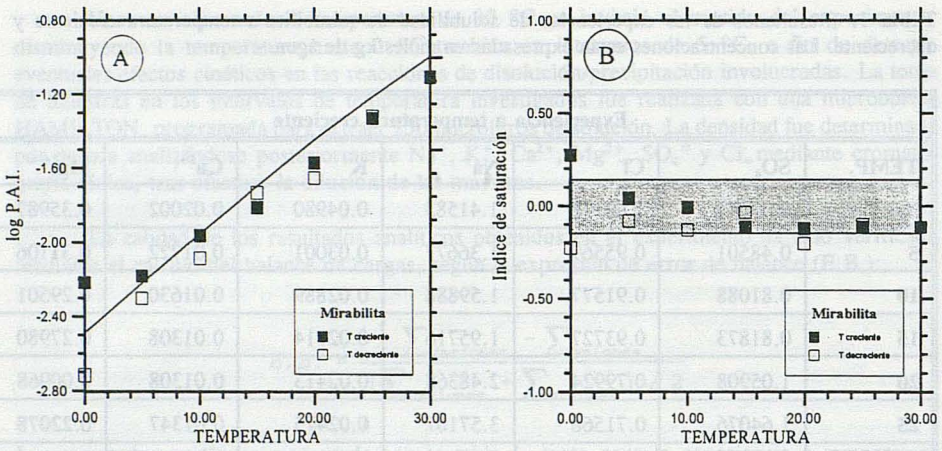


Figura 1. Productos de actividad iónica (P.A.I.) e índices de saturación respecto a la mirabilita calculados por el PHRQPITZ a partir de los datos experimentales, tanto a temperatura creciente como decreciente, presentados en la tabla 1. A.- Variación de log P.A.I. en el rango de temperaturas considerado. La curva de trazo continuo representa la variación con la temperatura de la constante de equilibrio de mirabilita considerada en el PHRQPITZ. B.- Valores de los índices de saturación respecto a ese mismo mineral. La franja oscura representa el intervalo de incertidumbre (I.S. = 0.0 ± 0.15) considerado para el cálculo de I.S. (ver texto).

Por ello, en este trabajo se considerará un rango de incertidumbre común a los resultados de I.S. de mirabilita, yeso y anhidrita de ± 0.15 unidades. O lo que es lo mismo, una situación de equilibrio solución-mineral no podrá determinarse con mayor precisión que la definida en el intervalo de I.S. = 0.0 ± 0.15 unidades.

4. Solubilidad de mirabilita y variación del quimismo de la solución

A partir de los datos presentados en la tabla 1 se han calculado con el PHRQPITZ los productos de actividad iónica de la mirabilita correspondientes a cada valor discreto de temperatura, representándose los valores obtenidos en la figura 1A. Como puede apreciarse los productos de actividad iónica calculados para este mineral, tanto a temperatura creciente como decreciente, son muy similares (salvo a 0 °C), definiendo una pauta creciente en el intervalo de temperatura considerado. Esta similitud de valores en los P.A.I. obtenidos a cada temperatura verifica el establecimiento de una situación de reversibilidad en la reacción mirabilita-solución y, por tanto, pueden considerarse como expresión del producto de solubilidad de ese mineral.

El producto de actividad iónica o producto de solubilidad, en forma logarítmica, obtenido a 25 °C (-1.327 y -1.316 en las experiencias a temperatura creciente y decreciente, respectivamente) concuerda bastante bien con el propuesto por Harvie & Weare (1980) y Harvie *et al.* (1984) de -1.228. Y además los obtenidos a las distintas temperaturas consideradas también coinciden aceptablemente bien con la variación de solubilidad de la mirabilita (representado por la curva a trazo continuo; fig. 1A) incluida en la base de datos del código PHRQPITZ. De hecho, los valores del índice de saturación respecto a la mirabilita, calculados a partir de los valores de

log K (T) de su base de datos y de los productos de actividad iónica experimentales, caen dentro del rango de I.S. 0.0 ± 0.15 (fig. 1B) para el que se considera una situación de equilibrio mineral-solución.

La única discrepancia apreciable se produce a 0 °C, con variaciones en los valores de log P.A.I. de -2.216 en la experiencia a temperatura creciente a -2.715 en la de temperatura decreciente), prácticamente equidistantes del producto de solubilidad considerado en el PHRQPITZ a esa temperatura (log K = -2.49). Pese a que esta temperatura se ha mantenido durante el doble de tiempo (48 horas) que el resto de valores discretos examinados en el experimento (ver Metodología), no se alcanza una situación de reversibilidad en la reacción, posiblemente a causa de efectos cinéticos asociados a la imposición de un valor tan extremo de temperatura en el dispositivo experimental. En todo caso, la ampliación del modelo HMW a temperaturas distintas de 25 °C realizada por Plummer et al. (1988) en el PHRQPITZ suministra resultados consistentes para la mirabilita en el resto del intervalo de temperaturas analizado.

El establecimiento de esta situación de reequilibrio mirabilita-solución al variar la temperatura tiene importantes consecuencias en el quimismo de la salmuera. En la figura 2A se han representado las concentraciones de $SO_4^=$ y Na^+, como componentes de mayor variabilidad en la experiencia. Sus rangos máximos de variación oscilan entre 0.24 moles/kg y 2.5 moles/kg para el $SO_4^=$ y entre 0.8 y 4.9 moles/kg para el Na^+ en el intervalo de temperatura considerado.

En la figura 2B se indica la evolución de los valores de la fuerza iónica de la solución y de la actividad del agua calculados por el PHRQPITZ a partir de los datos composicionales determinados para cada valor discreto de temperatura. Ambos parámetros pueden considerarse como expresión de las variaciones globales del quimismo de la solución durante el experimento, presentando un intervalo de variación de 1.9 a 8.1 molal para la fuerza iónica y de 0.966 a 0.896 para la actividad del agua.

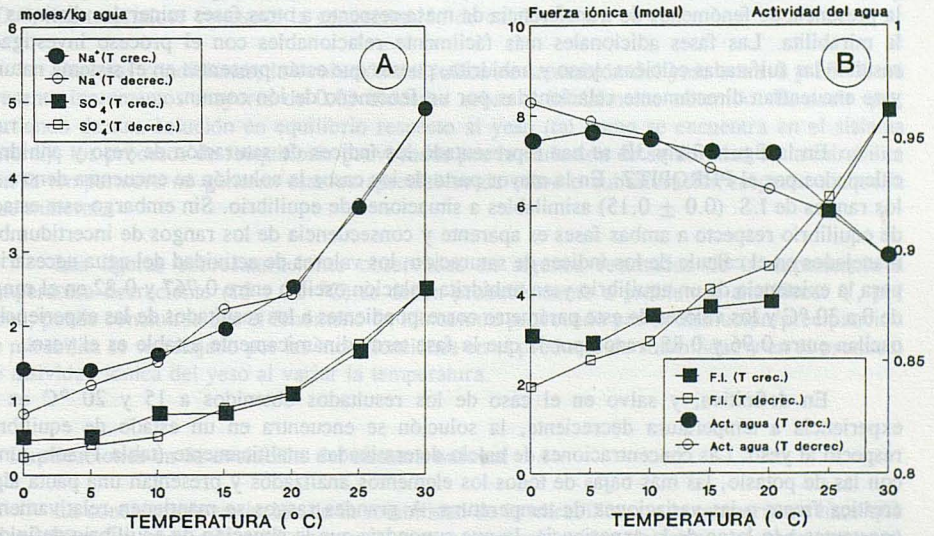


Figura 2. A.- Evolución de las concentraciones de Na^+ y $SO_4^=$ (en moles/kg de agua; tabla 1) respecto a la temperatura obtenidas en la experiencia de solubilidad de mirabilita. B.- Valores de fuerza iónica de la solución (molal) y actividad del agua calculadas por el PHRQPITZ a partir de los resultados experimentales (tabla 1).

Las concentraciones de $SO_4^=$ y Na^+ obtenidas tanto a temperatura creciente como decreciente son muy similares (tabla 1; fig. 2A), especialmente en el intervalo de 20 a 30 °C, existiendo pequeñas diferencias en los correspondientes al intervalo de 0 a 20 °C. Estas diferencias se ven ligeramente magnificadas en el caso de los valores de fuerza iónica y de actividad del agua (figura 2B) dado que se trata de parámetros en los que las diferencias de concentraciones se reflejan aditivamente.

En conjunto, las concentraciones de Na^+ y $SO_4^=$ siguen una pauta creciente conforme aumenta la temperatura si bien este crecimiento se hace especialmente marcado para el intervalo de 20 a 30 °C en el que la pauta de concentración de estos componentes incrementa de forma marcada su pendiente (figura 2A). Esta ruptura de pendiente se aprecia asimismo magnificada en la evolución de los valores de fuerza iónica y de actividad del agua (sobre todo en los resultados correspondientes al experimento a temperatura creciente) aunque, lógicamente, con pautas evolutivas opuestas (figura 2B).

Considerando, por tanto, las variaciones totales de las concentraciones de Na^+ y $SO_4^=$, de la fuerza iónica y de la actividad del agua en el rango de temperaturas examinado (0-30 °C) resulta evidente que la mayor parte de estas variaciones se produce en el intervalo de 20 a 30 °C. A grandes rasgos, más de un 50% de la variación observada en las concentraciones de estos elementos y en la actividad del agua tiene lugar en este intervalo, proporción que llega al 70 % en el caso de la fuerza iónica.

5. Estado de saturación de la solución respecto a yeso y anhidrita.

El proceso de reequilibrio mirabilita-solución al variar la temperatura provoca importantes modificaciones en el quimismo de la salmuera; modificaciones que a su vez pueden reflejarse en la presencia de fenómenos de transferencia de masa respecto a otras fases minerales distintas de la mirabilita. Las fases adicionales más fácilmente relacionables con el proceso investigado resultan las sulfatadas cálcicas, yeso y anhidrita, puesto que están presentes en el sistema natural y se encuentran directamente relacionadas por un fenómeno de ión común.

En la figura 3A y 3B se han representado los índices de saturación de yeso y anhidrita calculados por el PHRQPITZ. En la mayor parte de los casos la solución se encuentra dentro de los rangos de I.S. (0.0 ± 0.15) asimilables a situaciones de equilibrio. Sin embargo este estado de equilibrio respecto a ambas fases es aparente y consecuencia de los rangos de incertidumbre manejados en el cálculo de los índices de saturación: los valores de actividad del agua necesarios para la existencia de un equilibrio yeso-anhidrita-solución oscilan entre 0.767 y 0.82 en el rango de 0 a 30 °C y los valores de este parámetro correspondientes a los resultados de las experiencias oscilan entre 0.96 y 0.85, razón por la que la fase termodinámicamente estable es el yeso.

En definitiva, y salvo en el caso de los resultados obtenidos a 15 y 20 °C en la experiencia a temperatura decreciente, la solución se encuentra en un estado de equilibrio respecto al yeso. Las concentraciones de calcio determinadas analíticamente (tabla 1) son, junto con las de potasio, las más bajas de todos los elementos analizados y presentan una pauta algo errática frente a las variaciones de temperatura. A grandes rasgos se mantienen relativamente constantes a lo largo de la experiencia, lo que supondría que la situación de equilibrio definida para el yeso no implica la existencia de un proceso de transferencia de masa respecto a esa fase mineral como mecanismo que relaje los efectos inducidos por la disolución/precipitación de mirabilita.

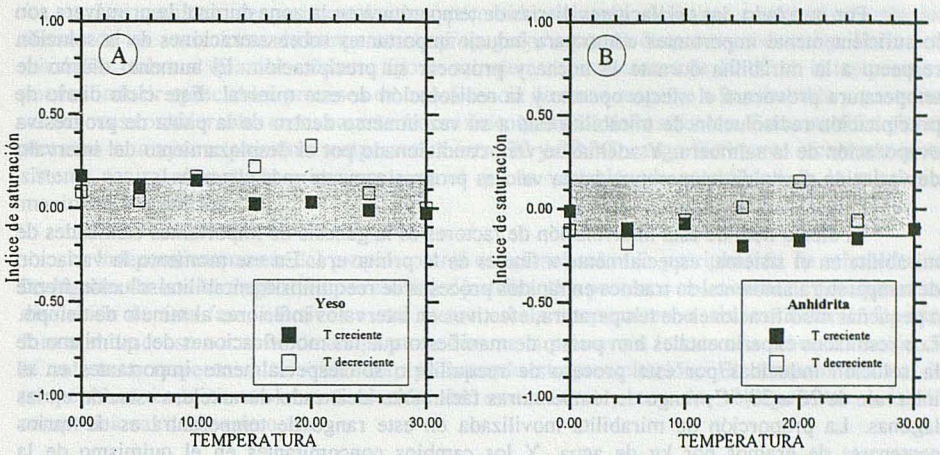


Figura 3. Índices de saturación de la solución respecto a (A) yeso y (B) anhidrita calculados por el PHRQPITZ a partir de los datos experimentales (tabla 1). La franja oscura representa el intervalo de incertidumbre (I.S. = 0.0 ± 0.15) considerado para el cálculo de I.S. (ver texto).

Para verificar esta situación y aislar este resultado de los valores analíticos de calcio se ha calculado teóricamente la variación del índice de saturación del yeso en una solución de características equivalentes a la del experimento a 0°C (temperatura creciente; tabla 1), equilibrada con mirabilita a temperaturas progresivamente crecientes (hasta 30 °C). El índice de saturación inicial de la solución respecto al yeso es de + 0.17 y, según el planteamiento del cálculo, los únicos componentes involucrados en procesos de transferencia de masa son Na^+ y $SO_4^=$, manteniéndose constante la concentración de calcio.

En estas condiciones, la evolución del índice de saturación del yeso entre 0 y 30 °C no presenta incrementos mayores de 0.03 unidades sobre el inicialmente obtenido. Es decir que, partiendo de una solución en equilibrio respecto al yeso (tal como se encuentra en el sistema natural), los procesos de reequilibrio provocados por la variación de solubilidad de la mirabilita con la temperatura no generan estados apreciables de sobre o subsaturación respecto a aquélla fase mineral.

Las ligeras sobresaturaciones observadas en algunos resultados de la experiencia a temperatura decreciente (15 y 20 °C) se deben probablemente a problemas analíticos. Y, por tanto, puede concluirse que el fenómeno de ión común provocado por la disolución/precipitación de mirabilita se ve relajado por las modificaciones en el producto de solubilidad y en el producto de actividad iónica del yeso al variar la temperatura.

6. Implicaciones en la evolución del sistema natural

La considerable variación de la solubilidad de la mirabilita con la temperatura hace que las modificaciones de este parámetro en el sistema natural incidan de forma determinante en el carácter estacional de su precipitación en las lagunas de Los Monegros, justificando las observaciones directas realizadas en este sistema (Pueyo, 1978-79; Pueyo e Inglés, 1987).

Por otro lado, las oscilaciones diarias de temperatura en la zona durante la primavera son lo suficientemente importantes como para inducir importantes sobresaturaciones de la solución respecto a la mirabilita durante la noche y provocar su precipitación. El aumento diurno de temperatura provocará el efecto opuesto y la redisolución de este mineral. Este ciclo diario de precipitación-redisolución de mirabilita está a su vez inmerso dentro de la pauta de progresiva evaporación de la salmuera. Y además se verá condicionado por el desplazamiento del intervalo de variación diaria de temperatura hacia valores progresivamente más elevados.

El efecto neto de esta interrelación de factores es la génesis de importantes cantidades de mirabilita en el sistema, especialmente a finales de la primavera. En ese momento la variación de temperatura ambiental se traduce en rápidos procesos de reequilibrio mirabilita-solución frente a pequeñas modificaciones de temperatura, efectivos en intervalos inferiores al minuto de tiempo. Los resultados experimentales han puesto de manifiesto que las modificaciones del quimismo de la solución inducidas por este proceso de reequilibrio son especialmente importantes en el intervalo de 20 a 30 °C, rango de temperaturas fácilmente alcanzado durante esa estación en las lagunas. La proporción de mirabilita movilizada en este rango de temperatura es de varios centenares de gramos por kg de agua. Y los cambios concomitantes en el quimismo de la solución llegan a duplicar los valores de su fuerza iónica y a modificar en casi 0.05 unidades la actividad del agua.

Estas drásticas modificaciones del quimismo de la salmuera no alteran, sin embargo, el estado de equilibrio respecto al yeso alcanzado en el transcurso del proceso de concentración por evaporación. Y por tanto no se producen efectos de transferencia de masa respecto a está fase mineral. Las concentraciones de calcio en la salmuera son considerablemente bajas (tabla 1) y el efecto de ión común provocado por las modificaciones del $SO_4^=$ en solución se ve compensado por la variación de la solubilidad del yeso y de su producto de actividad iónica con la temperatura.

Metodológicamente, la gran mayoría de los estudios realizados hasta el momento sobre la evolución geoquímica de sistemas endorreicos han utilizado aproximaciones basadas en la integración únicamente de datos composicionales procedentes de muestreos estacionales, anuales o plurianuales de las lagunas (Ordóñez *et al.*, 1991). Y, por otro lado, las aproximaciones más estrictamente fisicoquímicas al estudio de la secuencia de precipitación mineral por concentración evaporativa han sido normalmente planteadas en condiciones isotermas de 25 °C. La utilización de estos planteamientos (obviando explícitamente la temperatura o considerando un valor constante de 25 °C) impedirá una precisa evaluación de las condiciones mineralogenéticas en medios como los analizados en este trabajo, en los que los efectos de las variaciones de temperatura, aislados de los propiamente evaporativos, pueden llegar a ser importantes.

7. Conclusiones

Se ha determinado experimentalmente la variación de solubilidad de la mirabilita en un rango de 0 a 30 °C, utilizando la propia salmuera natural de una de las lagunas de Los Monegros. El cálculo preciso de los productos de actividad iónica en esta solución altamente concentrada y de composición relativamente compleja se ha realizado mediante el código PHRQPITZ (Plummer et al., 1988) con una nueva propuesta de ampliación del modelo HMW (Harvie & Weare, 1980; Harvie *et al.*, 1984) a temperaturas distintas de 25 °C. El producto de solubilidad obtenido a 25 °C es de -1.321 $\pm$ 0.006, ligeramente superior al considerado en el modelo HMW de -1.228; y los resultados a las demás temperaturas se ajustan a los definidos por la ampliación de este modelo incluida en el PHRQPITZ.

Los resultados experimentales han permitido, además, delimitar cuantitativamente los efectos de las variaciones de temperatura, propagados a través de procesos de reequilibrio con mirabilita, sobre el quimismo de este tipo de salmueras. Oscilaciones de 10 °C en la temperatura pueden inducir apreciables procesos de transferencia de masa respecto a la mirabilita y provocar variaciones considerables en las características fisicoquímicas de la solución. Estas modificaciones alcanzan su mayor expresión a temperaturas entre 20 y 30 °C, rango fácilmente alcanzado en el sistema natural durante la primavera y en el que la variación de solubilidad de la mirabilita maximiza sus efectos.

Independientemente del propio control estacional sobre la precipitación de mirabilita en las lagunas de Los Monegros, estos resultados justifican los procesos observados de precipitación-disolución de esta fase, efectivos a una menor escala temporal en el sistema natural. De esta forma el descenso nocturno de temperatura durante la primavera induce una sobresaturación de la salmuera respecto a la mirabilita, provocando su precipitación. El aumento diurno de temperatura condiciona la redisolución de este mineral, con la subsiguiente variación composicional de la salmuera.

Los efectos de este proceso se hacen especialmente evidentes cuando el sistema alcanza un estado de concentración evaporativa importante, caracterizado por la génesis de considerables cantidades de mirabilita en el medio. En esas condiciones el proceso dispone de la suficiente cantidad de mirabilita como para evolucionar, incluso durante las etapas de incremento de temperatura, a través de rápidos procesos de reequilibrio respecto a esta fase. Y además, la temperatura ambiente en esa etapa de evolución de la laguna se moverá fundamentalmente en el intervalo de 20 a 30 °C, removilizando importantes cantidades de mirabilita (varios centenares de gramos por kg de solución) y amplificando los efectos composicionales sobre la salmuera remanente.

Estas variaciones en el quimismo de la solución no alteran, sin embargo, su estado de equilibrio respecto al yeso. Este equilibrio se mantiene invariablemente pese a los cambios de temperatura y a los efectos de ión común, sin necesidad de que se produzcan procesos de reequilibrio yeso-solución en el sistema.

En cualquier caso, la incidencia de las oscilaciones de temperatura ambiente en este tipo de sistemas, a través de la génesis de sales especialmente sensibles a este parámetro, son lo suficientemente importantes como para ser sistemáticamente ignorados dentro de las metodologías usuales de estudio de su evolución geoquímica.

Bibliografía

Bischoff, J.L.; Herbst, D.B. & Rosenbauer, R.J. (1991). Gaylussite formation at Mono Lake, California. *Geochim. Cosmochim. Acta*, 55, 1743-1747.

Bischoff, J.L.; Stine, S.; Rosenbauer, R.J.; Fitzpatrick, J.A. & Stafford Jr., T.W. (1993). Ikaite precipitation by mixing of shoreline springs and lake water, Mono Lake, California, USA. *Geochim. Cosmochim. Acta, 57*, 3855-3865.

Fisher, R.S. & Kreitler, C.W. (1987). Geochemistry and hydrodynamics of deep-basin brines Palo Duro Basin, Texas, USA. *Applied Geochemistry*, 2, 459-476.

Garcés, I.; Tena, J.M.; Auqué, L.F.; Gimeno, M.J. y Mandado, J. (1991). Variación de los índices de saturación en función del cálculo de coeficientes de actividad. Su aplicación a las fases mineralógicas de las lagunas de Monegros (Zaragoza, España). *Estudios Geológicos*, 47, 305-315.

Garcés, I.; Tena, J.M.; Auqué, L.F.; Mandado, J. y Gimeno, M.J. (1992). Evolución geoquímica de las salmueras de las lagunas de Monegros (Zaragoza, España) y análisis del equilibrio de la halita. *Estudios Geológicos*, 48, 101-110.

Greenberg, J.P. & Moller, N. (1989). The prediction of mineral solubilities in natural waters: A chemical equilibrium model for the Na-K-Ca-Cl-SO_4-H_2O system to high concentration and from 0 to 250 °C. *Geochim. Cosmochim. Acta*, 53, 2503-2518.

Harvie, C.E. & Weare, J.H. (1980). The prediction of mineral solubilities in natural waters: the Na-K-Mg-Ca-Cl-SO_4-H_2O system from zero to high concentration at 25 °C. *Geochim. Cosmochim. Acta,* 44, 981-997.

Harvie, C.E.; Moller, N. & Weare J.H. (1984). The prediction of mineral solubilities in natural waters: the Na-K-Mg-Ca-H-Cl-SO_4-OH-HCO_3-CO_3-CO_2-H_2O system to high ionic strengths at 25 °C. *Geochim. Cosmochim. Acta*, 48, 723-751.

Jenne, E.A.; Ball, J.W.; Burchard, J.M.; Vivit, D.V. & Barks, J.H. (1980). Geochemical modeling: apparent solubility controls on Ba, Zn, Cd, Pb and F in waters of the Missouri Tri State mining area. In: D.D. Hemphill (ed.), *Trace substances in environmental health.* XIV. University of Missouri, Columbia, Mo., 353-361.

Langmuir, D. & Melchior, D. (1985). The geochemistry of Ca, Sr, Ba and Ra in some deep brines from the Palo Duro Basin, Texas, USA. *Geochim. Cosmochim. Acta*, 49, 2423-2432.

Mingarro, F.; Ordóñez, S. López de Azcona, M.C. y García del Cura, M. de los A. (1981). Sedimentoquímica de las lagunas de Los Monegros y su sistema geológico. *Bol. Geol. Min.*, T. XCII-III, 171-195.

Moller, N. (1988). The prediction of mineral solubilities in natural waters: A chemical equilibrium model for the Na-Ca-Cl-SO_4-H_2O system, to high temperature and concentration. *Geochim. Cosmochim. Acta*, 52, 821-837.

Ordoñez, S.; Sánchez del Moral, S. y García del Cura, M.A. (1991). Modelización de la hidroquímica de una laguna tipo playa (Cl^--$SO_4^=$-Mg^{2+}-Na^+): La Laguna Grande de Quero (Toledo). *Est. Geológicos*, 47, 207-219.

Pitzer, K.S. (1973). Thermodynamics of electrolytes. I: Theoretical basis and general equations. *J. Phys. Chem.*, 77, 268-277.

Pitzer, K.S. (1979). *Activity coefficients in Electrolyte Solutions*, Chap. 7. CRC Press, Boca Raton, FL.

Plummer, L.N.; Parkhurst, D.L.; Fleming, G.W. and Dunkle, S.A. (1988). A computer program incorporating Pitzer's equations for calculation of geochemical reactions in brines. *U.S. Geological Survey, Water Resources Investigations Report 88-4153.*

Pueyo, J.J. (1978-79). La precipitación evaporítica actual en las lagunas saladas del área de Bujaraloz, Sástago, Caspe, Alcañiz y Calanda (provincias de Zaragoza y Teruel). *Rev. Inst. Inv. Geol. Dip. Prov. Barcelona*, 33, 5-56.

Pueyo, J.J. (1980). Procesos diagenéticos observados en las lagunas tipo playa de la zona de Bujaraloz-Alcañiz (provincias de Zaragoza y Teruel). *Rev. Inst. Inv. Geol. Dip. Prov. Barcelona*, 34, 195-207.

Pueyo, J.J. e Inglés-Urpinell, M. (1987). Substrate mineralogy, pore brine compositions and diagenetic processes in the playa lakes of Los Monegros and Bajo Aragon, Spain. In: (R, Rodríguez-Clemente and Y. Tardy, eds.) *Geochemistry and Mineral Formation in the Earth Surface*, 351-372.

Pueyo, J.J. y De la Peña, J.A. (1991). Los lagos salinos españoles. Sedimentología, hidroquímica y diagénesis. In: (J.J. Pueyo, coord.) *Génesis de Formaciones Eavporíticas. Modelos Andinos e Ibéricos*, 163-192.

Rezagui, M. (1993). *Dynamique des sels dans les eaux et dans les plates halophytes (Salicornia L.) dans deux régions arides (Algérie et Espagne).* Thèse Magistère. Université de Annaba (Algeria). 71 pp.

Smith, G.I.; Friedman, I. & McLaughlin, R.J. (1987). Studies of Quaternary saline lakes. III. Mineral, chemical and isotopic evidence of salt solution and crystallization processes in Owens Lake, California, 1969-1971. *Geochim. Cosmochim. Acta*, 51, 811-827.

Vallès, V.; Rezagui, M. ; Auqué, L.; Semadi, A. y López, P. (1994). Geochimie des sols salees de deux zones arides du pourtour Mediterraneen. II. Geochimie des sels de deux lagunes salees du Bas Aragon en Espagne. *Arid Soil Res. (enviado)*