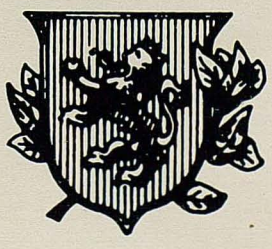
REVISTA
DE LA

ACADEMIA DE CIENCIAS

Exactas
Físicas
Químicas y
Naturales

DE
ZARAGOZA



Serie 2.ª
Volumen 48

1993

INDICE DE MATERIAS

Rev. Academia de Ciencias. Zaragoza. 48 (1993)

FOLIATIONS, GROUPOIDS AND GROTHENDIECK ETENDUES

Ieke Moerdijk

Mathematical Institute Utrecht University

Budapestlaan 6 , P.O. box 80.010, 3508 TA Utrecht,

The Netherlands

Prologue

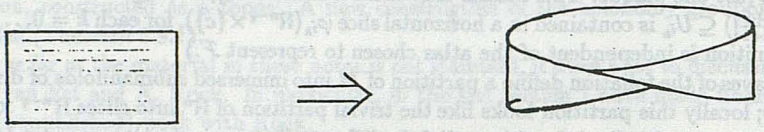
This is the text of five lectures given at the University of Zaragoza in January 1993. The purpose of the lectures was to give an introduction to the relation between foliations and Grothendieck topoi.

In the study of foliations, a central rôle is played by the so-called space of leaves of the foliation. It is immediately clear from elementary examples that the naively constructed quotient space $M/\sim$ of leaves of a foliation on a manifold M does in general contain very little information. One is thus led to developing a more general notion of "space" which can capture the topological and differential-geometric properties of the space of leaves, and many proposals have been made in this direction. (A brief list occurs in Section 5 below.) One of the first such proposals was made by Grothendieck (SGA4, vol. 1, p. 485), who associated to each foliation $\mathcal{F}$ of a manifold M first a so-called local equivalence relation $r_{\mathcal{F}}$ on M and then a topos $\mathrm{Sh}(M, r_{\mathcal{F}})$ of $r_{\mathcal{F}}$-invariant sheaves. This topos is of a special kind, a so-called differentiable (smooth) étendue. A closely related topos is the topos of holonomy-invariant sheaves on the manifold M, which I will denote by $M/\mathcal{F}$; this is also a smooth étendue. The more specific goal of these lectures is to illustrate that this topos $M/\mathcal{F}$ serves as an excellent model for the "space of leaves" of the foliated manifold $(M, \mathcal{F})$. Many invariants (such as the cohomology groups and the (etale) homotopy groups) are immediately naturally defined for $M/\mathcal{F}$, as part of the general theory of topoi. It will (very briefly) be discussed how these invariants relate to some transversal invariants studied in the literature, e.g. the cohomology of basic forms. As part of general topos theory, one also obtains a natural notion of (smooth) map $M/\mathcal{F} \to M'/\mathcal{F}'$ between two such spaces of leaves. It will be shown that the localization theorem from Moerdijk (1988) implies that such maps between topoi $M/\mathcal{F} \to M'/\mathcal{F}'$ correspond exactly to the maps between "leaf spaces" utilized in the work of Connes, Skandalis, and others.

The level of these lectures is rather introductory. In particular, the first lecture discusses the definition and some well-known examples of foliated manifolds. This material can be

Depending on α, each leaf winds around the torus a fixed finite number of times, or infinitely often. In the latter case the leaves are (as always immersed but) not embedded. Each leaf is dense. The quotient topology on the space of leaves is the trivial (indiscrete) topology in this case.

1.3.3 The Möbius band. Foliate the square $I^2 = [0,1]^2$ by horizontal lines. Next construct the Möbius band by identifying the two opposite sides while reversing the orientation:



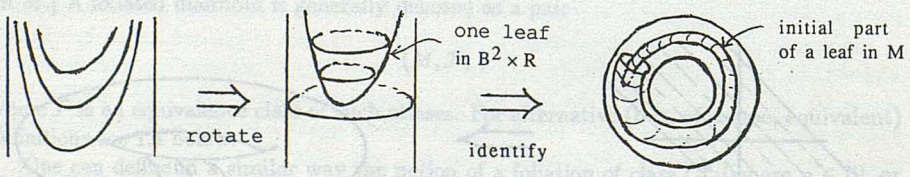
This gives a foliation on the Möbius band whose leaves are all circles. There is a central leaf going around once, while the other leaves are "double coverings" of this central one. (In fact this is a foliation of a manifold with boundary. Of course there is a similar construction of an "infinite" foliated Möbius band obtained from $I \times \mathbf{R}$ by identifying $(0,x)$ and $(1,-x)$.)

1.3.4 Suspension. The previous example is the suspension of the diffeomorphism $t \mapsto 1-t : [0,1] \to [0,1]$ (or $x \mapsto -x : \mathbf{R} \to \mathbf{R}$ for the infinite band). More generally, if $f : M \to M$ is any diffeomorphism of a manifold M, the trivial foliation on $M \times \mathbf{R}$ whose leaves are the vertical lines $\{x\} \times \mathbf{R}$ gives a foliation on the quotient space

$$M_f = (M \times \mathbf{R})/\sim$$

obtained by identifying $(x, t+1)$ with $(f(x), t)$. Write $q : M \times \mathbf{R} \to M_f$ for the quotient map (a covering projection). Inside M_f there is an embedded copy $\overline{M} = q(M \times \{0\})$ of M, which is *transverse* to each leaf. The leaf through $q(x,0)$ "returns" to $\overline{M}$ in $q(f(x),0)$. So f is recovered by following the leaves until one first gets back to $\overline{M}$ (Poincaré map of "first return").

1.3.5 The Reeb foliation. This is a foliation of the solid torus $M = B^2 \times S^1$ (where B^2 is the 2-ball and S^1 the 1-sphere). To describe this foliation, think of M as constructed from an infinite solid cylinder $B^2 \times \mathbf{R}$ by identifying (x,t) and $(x,t+1)$. Now foliate this infinite solid cylinder by a foliation whose leaves are the boundary $S^1 \times \mathbf{R}$ and a collection of infinitely deep salad bowls piled onto (into) each other, and filling up the interior of the solid cylinder. (The picture on the left is a segment $[-1,1] \times \mathbf{R}$.)



Note that each of the leaves in the interior of M is a copy of $\mathbf{R}^2$ which accumalates on the boundary of M.

The 3-sphere S^3 can be constructed by glueing two solid tori together along their boundaries. Thus one obtains a codimension 1 C^∞-foliation of S^3 from two copies of the

Reeb foliation on the solid torus. By Haefliger's theorem [Haefliger(1958)] there exists no analytic foliation on S^3 of codimension 1. Any compact oriented connected 3-manifold can be obtained from S^3 by cutting out solid tori and sewing them back in with a twist [Lickorish(1962), (1965)]. This can be used, together with the Reeb foliation, to construct a foliation of codimension 1 for any such 3-manifold.

1.3.6 Flat fiber bundles. Let $p : E \to B$ be a smooth fiber bundle, with fiber F. Then in particular p is a submersion, so E carries a foliation whose leaves are (the components of) the fibers of p, as in 1.3.1. A more interesting foliation, whose leaves are "horizontal" (rather than vertical) parts of E, can be constructed if the bundle is *flat*. This means that its structure group G is discrete. Indeed, suppose $\varphi_i : U_i \times F \xrightarrow{\sim} p^{-1}(U_i) \subseteq E$ are the local charts for the bundle, for an open cover $B = \bigcup U_i$. Write $\{\varphi_{ij} : U_i \cap U_j \to G\}$ for the associated cocycle into the discrete structure group G. By choosing the U_i suitably, we may assume that the intersections $U_i \cap U_j$ are connected, so that each φ_{ij} is constant. Now define a trivial foliation on each $p^{-1}(U_i) \subseteq E$ with as leaves the horizontal slices $\varphi_i(U_i \times \{x\})$ for $x \in F$. These fit together to give a foliation on E. Its holonomy (§2) reflects the cohomology class in $H^1(X, G)$ of the cocycle $\{\varphi_{ij}\}$.

1.4 Alternative definition of foliations. We mention three different, but equivalent, ways of defining foliations.

1.4.1 Submersion form. A C^∞-foliation on a C^∞-manifold M, of codimension q, can be given by an open covering $M = \bigcup U_i$ and submersions $\psi_i : U_i \to \mathbf{R}^q$ with the property that for each intersection $U_i \cap U_j$ there exists a C^∞-diffeomorphism h_{ij} making the diagram

$$\begin{array}{ccc} & U_i \cap U_j & \\ {\scriptstyle \psi_i}\swarrow & & \searrow{\scriptstyle \psi_j} \\ \psi_i(U_j) & \xrightarrow{\quad h_{ij} \quad} & \psi_j(U_i) \end{array}$$

commute. This definition is related to the earlier one (1.1) by defining $\psi_i = \pi_2 \circ \varphi_i^{-1} : U_i \to \mathbf{R}^{n-q} \times \mathbf{R}^q \to \mathbf{R}^q$.

1.4.2 Integrable subbundle form. A C^∞-foliation on M of codimension q defines an $(n-q)$-dimensional subbundle $T_{\text{leaf}}(M)$ of the tangent bundle TM, consisting of those tangent vectors which are tangent to the leaves. This subbundle is ***integrable***, in the sense that if X, Y are sections of $T_{\text{leaf}}(M)$ on an open set $U \subseteq M$ then their Lie-bracket $[X, Y]$ lies again in the subbundle $T_{\text{leaf}}(M)$. Conversely, by the theorem of Frobenius (1877), an integrable subbundle $E \subseteq TM$ (of codimension q) defines a unique foliation on M for which $E = T_{\text{leaf}}(M)$.

1.4.3 Local equivalence relations. (Grothendieck-Verdier, SGA4, vol.1, p. 485.) For any space M there is a presheaf $\mathcal{E}$ on M, defined for each open set $U \subseteq M$ by

$$\mathcal{E}(U) = \{R \mid R \subseteq U \times U \text{ is an equivalence relation on } U\},$$

with the evident restriction maps $\mathcal{E}(U) \to \mathcal{E}(V)$ for open sets U, V with $V \subseteq U$. This presheaf is not a sheaf, and we write $\tilde{\mathcal{E}}$ for its associated sheaf. A *local equivalence relation*

$\underline{\mathrm{r}}$ on M is by definition a global section of $\tilde{\mathcal{E}}$. Such an l.e.r. is given by an *atlas*, i.e. an open cover $M = \bigcup U_i$, together with equivalence relations $R_i \subseteq U_i \times U_i$, which are locally compatible; this means that for each $x \in U_i \cap U_j$ there is a small neighborhood $W_x \subseteq U_i \cap U_j$ so that R_i and R_j restrict to the same equivalence relation on W_x. Each foliation gives rise to a local equivalence relation (with the plaques as equivalence classes). Conversely, by Godement's theorem [Serre(1992), p. 92], an l.e.r. $\underline{\mathrm{r}}$ defines a foliation of codimension Q when there exists an atlas $\{(U_i, R_i)\}$ for $\underline{\mathrm{r}}$ for which each $R_i \subseteq U_i \times U_i$ is a closed submanifold of codimension q for which the two projections $R_i \to U_i$ are both submersions.

1.5 Transversal structure. An important aspect of the theory of foliations is the study of the "tranversal structure". The cohomology of basic differential forms (see §5) is a typical example. The most naive approach would be to consider the quotient space $M/\sim$ obtained from M by identifying any two points which are in the same leaf. The elementary examples given above already illustrate that this naively constructed space of leaves in general does not retain much of the structure of the foliation. An underlying theme for these lectures is the problem of constructing a more sophisticated quotient "space" of leaves.

1.6 References for this lecture. Apart from the approach by Grothendieck using local equivalence relations (1.4.3), the material in this section is completely standard and can be found in almost any exposition of the theory of foliations; e.g. Camacho-Neto (1985), Fuks (1982), Godbillon (1991), Hector-Hirsch (1981), Lawson (1977), Molino (1988), Tondeur (1988), and many others.

2 Groupoids and holonomy

The rest of these lectures presupposes some knowledge of category theory, as contained e.g. in the first chapters of [MacLane (1971)]. We begin by reviewing the basic definitions and fixing the notation.

2.1 Categories. A category C is given by a collection C_0 of *objects*, a collection C_1 of *arrows*, and four structure maps:

$$C_1 \times_{C_0} C_1 \xrightarrow{m} C_1 \underset{d_1}{\overset{d_0}{\underset{\longleftarrow}{\overset{\longrightarrow}{i}}}} C_0 .$$

The maps d_0 and d_1 give for each arrow $\alpha \in C_1$ its *domain* $d_0(f)$ and its *codomain* $d_1(f)$. The map m is defined for any pair of arrows α, β with $d_0\alpha = d_1\beta$, and assigns to this pair the *composition* $m(\alpha, \beta)$, also denoted $\alpha \circ \beta$. Finally, the map i assigns to each object $x \in C_0$ the identity arrow at x, denoted $i(x)$ (or id_x, or 1_x). These maps must satisfy the well-known identities

$$\begin{array}{ll} d_0 i(x) = x = d_1 i(x) & (\alpha \circ \beta) \circ \gamma = \alpha \circ (\beta \circ \gamma) \\ d_0(\alpha \circ \beta) = d_0\beta & \alpha \circ i(d_0\alpha) = \alpha \\ d_1(\alpha \circ \beta) = d_1\alpha & i(d_1\alpha) \circ \alpha = \alpha. \end{array}$$

One writes $\alpha : x \to y$ to denote that $\alpha \in C_1$ is an arrow *from* $x = d_0(\alpha)$ *to* $y = d_1(\alpha)$.

For example, the category **Sets** has as objects all sets, and as arrows all functions between sets. It is a very large category. As another example, the "simplicial model category" Δ has as objects the finite ordered sets

$$[n] = \{0, \ldots, n\} \qquad (\text{all } n \geq 0);$$

the arrows $[n] \to [m]$ of the category Δ are the monotone functions $\alpha : [n] \to [m]$. As a final example, recall that a *monoid* is a set M equipped with an associative operation $\mu : M \times M \to M$ which has a 2-sided unit $e \in M$. Such a monoid can be viewed as a category with just one object $*$: the arrows $* \to *$ are the elements of M, with e as identity arrow and with composition given by μ.

For each category C there is a "dual" category C^{op}, called the *opposite category* of C. It has the same objects and arrows as C, but the domain and codomain are interchanged: an arrow $\alpha : x \to y$ in C is an arrow $\alpha : y \to x$ in C^{op}. The identities and composition of C^{op} are the same as those of C.

2.2 Groupoids. An arrow $\alpha : x \to y$ in a category C is said to be an *isomorphism* if there exists an arrow $\beta : y \to x$ so that $\alpha \circ \beta = i(y)$ and $\beta \circ \alpha = i(x)$. The objects x and y are then said to be isomorphic. Given α, the arrow β is unique (if it exists) and is denoted α^{-1}; it is called the *inverse* of α. A *groupoid* is a category C in which each arrow is an isomorphism. For example, the category of finite sets and bijections is a groupoid. Another example is the *fundamental groupoid* of a topological space. More generally, if X is a space and $S \subseteq X$ is a set of points, one can construct the fundamental groupoid relative to the set S of "base-points"

$$\Pi(S, X);$$

its objects are the points of S, and for $x, y \in S$ an arrow $x \to y$ is a homotopy class $[\alpha]$ of paths $\alpha : [0,1] \to X$ from $x = \alpha(0)$ to $y = \alpha(1)$. Composition is defined just as for the fundamental group of X. Each group G also gives an example of a groupoid, with just one object and the elements of the group as arrows (as in the case of monoids above).

2.3 Functors. A *functor* between two categories $F : C \to D$ is given by two operations on objects and arrows, both denoted F:

$$F : C_0 \to D_0, \quad F : C_1 \to D_1,$$

which respect the structure maps of the category; i.e.,

$$\begin{array}{ll} F(i(x)) = i(F(x)) & (\text{each } x \in C_0), \\ F(d_k\alpha) = d_k F(\alpha) & (k = 0, 1, \text{ each } \alpha \in C_1), \\ F(\alpha \circ \beta) = F(\alpha) \circ F(\beta) & (\text{all } \alpha, \beta \in C_1 \text{ with } d_0\alpha = d_1\beta). \end{array}$$

For example, when groups G and H are viewed as one-object categories, a functor $G \to H$ is the same thing as a homomorphism. As another example, a *simplicial set* is by definition the same thing as a functor $\Delta^{op} \to \text{Sets}$.

For two functors $F, G : C \to D$, a *natural transformation* $\tau : F \to G$ is a family of arrows in D,

$$\tau_x : F(x) \to G(x), \quad \text{for } x \in C_0,$$

such that for each arrow $\alpha : x \to y$ in C the square

$$\begin{array}{ccc} F(x) & \xrightarrow{\tau_x} & G(x) \\ {\scriptstyle F(\alpha)}\downarrow & & \downarrow{\scriptstyle G(\alpha)} \\ F(y) & \xrightarrow{\tau_y} & G(y) \end{array}$$

commutes. Such natural transformations between functors can be composed, and one obtains a category $\mathrm{Hom}(C, D)$ with functors as objects and natural transformations as arrows. A *natural isomorphism* is a natural transformation $\tau : F \to G$ with the property that each $\tau_x : F(x) \to G(x)$ is an isomorphism in the category D. This is equivalent to saying that τ is an isomorphism in the functor category $\mathrm{Hom}(C, D)$.

A functor between groupoids will also be called a homomorphism. The category $\mathrm{Hom}(C, D)$ is again a groupoid when D is a groupoid.

2.4 Equivalence of categories. A functor $F : C \to D$ is said to be a (weak) equivalence if

(i) F is *essentially surjective*: for each object y of D there exists an object x of C and an isomorphism $\alpha : i(x) \tilde{\to} y$;

(ii) F is *full and faithful*: for any two objects x, x' of C, the functor F induces a bijection

$$F : C(x, x') \to D(F(x), F(x')) .$$

Here $C(x, x')$ denotes the set of all arrows from x to x'. If such an equivalence exists, C and D are said to be *equivalent categories*. (This is a symmetric notion, see [MacLane (1971)].)

2.5 Topological categories and groupoids. A *topological category* is a category C where C_0 and C_1 are sets equipped with a topology, such that the four structure maps are continuous w.r.t. these two topologies. Such a topological category C is called a *topological groupoid* if there is a continuous operation

$$C_1 \to C_1, \qquad \alpha \mapsto \alpha^{-1} \tag{2.5.1}$$

assigning to each arrow α its inverse. For example, the fundamental groupoid $\Pi(S, X)$ mentioned above is a topological groupoid in a natural way. Also, each topological group provides an example of a topological groupoid. If the spaces C_0 and C_1 are C^∞-manifolds (we will generally assume that C_0 is Hausdorff, but C_1 need not be) and $d_0, d_1 : C_1 \rightrightarrows C_0$ are differentiable submersions then the fibered product $C_1 \times_{C_0} C_1$ is again a C^∞-manifold. If moreover the composition $m : C_1 \times_{C_0} C_1 \to C_1$ and the map $i : C_0 \to C_1$ are differentiable, then C is called a *differentiable category*. If moreover C is a groupoid for which the operation $\alpha \mapsto \alpha^{-1}$ of 2.5.1 is differentiable then C is called a *differentiable* (or *smooth*) *groupoid*. For example, each Lie group is a smooth groupoid. By a *homomorphism* of topological (or differentiable) groupoids $F : C \to D$ we will mean a functor which is continuous (differentiable) w.r.t. the given topologies (manifold structures) on C and D.

2.6 Haefliger's groupoid Γ^q (for $q \geq 0$ an integer). This is the groupoid with $\mathbf{R}^q$ as space of objects. For two points x and y in $\mathbf{R}^q$, the arrows $x \to y$ of Γ^q are germs of diffeomorphisms. More explicitly, consider diffeomorphisms $\alpha : U \tilde{\to} V$ between open neighbourhoods U of x and V of y, such that $\alpha(x) = y$. Call two such $\alpha : U \to V$ and $\beta : U' \to V'$ equivalent if they agree on a neighbourhood $W \subseteq U \cap U'$ of x. The equivalence class of $\alpha : U \to V$ is called the *germ* of α at x, and denoted $\mathrm{germ}_x(\alpha)$, or simply α_x. These germs, for all $x, y \in \mathbf{R}^q$, are the arrows of Γ^q. The set of germs can be equipped with a natural (sheaf) topology, making Γ^q into a differentiable groupoid, with the additional property that the domain and codomain maps are etale (local diffeomorphisms).

N.B. For the spaces of objects and of arrows of Γ^q one generally does not write $(\Gamma^q)_0$ and $(\Gamma^q)_1$ but $\mathbf{R}^q$ and Γ^q.

2.7 Etale groupoids. A topological groupoid G is said to be *etale* if the domain and codomain maps d_0, d_1: $G_1 \rightrightarrows G_0$ are both etale maps, i.e. local homeomorphisms. If G is differentiable, G is called etale if d_0 and d_1 are local diffeomorphisms. For example, the groupoid Γ^q is etale, as are the following two types of examples.

2.8 S-atlases. [Van Est (1984)] Let G be an etale topological groupoid, and consider an arrow $g : x \to y$ of G. Since $d_0, d_1 : G_1 \to G_0$ are assumed etale, there is a small neighbourhood U_g of g such that d_0 and d_1 restrict to homeomorphisms $U_g \tilde{\to} V_x$ and $U_g \tilde{\to} W_y$ for suitable neighbourhoods V_x resp. W_y of x and y. The map $d_1 \circ d_0^{-1} : V_x \to U_g \to W_y$ sends x to y, and its germ only depends on g. We write $\tilde{g} := \mathrm{germ}_x(d_1 \circ d_0^{-1})$. This representation of arrows g of G by germs of diffeomorphisms need not be faithful (e.g. take G a Lie group). By definition, an S-atlas is a differentiable groupoid with the property that the germ-representation is faithful: for any two arrows $g, h : x \to y$, if the germs $\tilde{g}$ and $\tilde{h}$ are identical, then $g = h$. For example, Γ^q is an S-atlas. Another example is the etale holonomy groupoid $\mathrm{Hol}_T(M, \mathcal{F})$ of a foliation, discussed in 2.11 below.

2.9 QF-varieties. [Pradines, Wouafa-Kamga (1979)]. A QF-variety is represented by a surjective map $p : M \to S$, where M is a C^∞-manifold and S is just a set, such that several conditions are satisfied. One of these is that for any two points $x, y \in M$ with $p(x) = p(y)$, there is a diffeomorphism $f : V_x \tilde{\to} V_y$, between neighbourhoods of x and y, such that $p \circ f = p$. (The other conditions need not concern us here.) Such a QF-variety gives rise to a differentiable groupoid with M as space of objects, and with arrows $x \to y$ all germs of diffeomorphisms f as above. This space of germs can be given the sheaf topology, so that the resulting groupoid is etale; in fact it is an S-atlas. [A QF-variety is actually given as an equivalence class of such maps $f : M \to S$; equivalent such maps give weakly equivalent differentiable groupoids (2.12 below), hence the same étendue (3.2).]

2.10 Monodromy of a foliation. Let $(M, \mathcal{F})$ be a foliated manifold. We construct a topological groupoid, the monodromy groupoid of the foliation,

$$\mathrm{Mon}(M, \mathcal{F}),$$

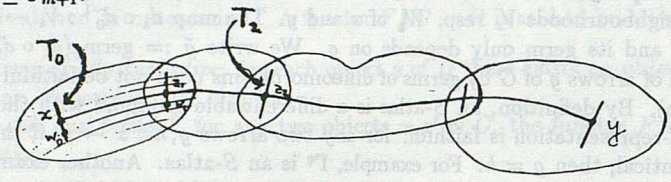
with the manifold M as space of objects. For points $x, y \in M$, there are arrows $x \to y$ in $\mathrm{Mon}(M, \mathcal{F})$ only if x and y belong to the same leaf. Writing L_x for this leaf, an arrow $x \to y$ is a homotopy class of paths from x to y in L_x. (The paths as well as the homotopies should be in L_x; and as usual the homotopies should leave the endpoints fixed.) This set of arrows carries a natural topology, derived from the compact-open topology on the function space M^I, and making $\mathrm{Mon}(M, \mathcal{F})$ into a topological groupoid (in fact even a differentiable groupoid). For details of the construction, see [Philips (1987)], [Kock-Moerdijk (1991), (in preparation)].

2.11 Holonomy of a foliation. Let $(M, \mathcal{F})$ be a foliated manifold, as before. The *holonomy groupoid* of $(M, \mathcal{F})$, denoted

$$\mathrm{Hol}(M, \mathcal{F}),$$

is a suitable quotient of $\mathrm{Mon}(M, \mathcal{F})$. It was introduced by [Winkelnkemper (1983)], who calls it the graph of the foliation, and discussed further e.g. in [Connes (1982)] and [Haefliger (1984)]. The definition can be outlined as follows. The space of objects of $\mathrm{Hol}(M, \mathcal{F})$ is again the underlying manifold M. For two points x and y in M, there are arrows $x \to y$ in

the holonomy groupoid only if x and y are in the same leaf, and they are again equivalence classes of paths inside a leaf, just as for monodromy. However, the equivalence relation is different. Consider a path $\alpha : [0,1] \to L \subseteq M$ into the leaf L of $x = \alpha(0)$ and $y = \alpha(1)$. As in section 1.2 , we can find a chain $U_0, \ldots, U_m$ of coordinate neighbourhoods for the foliation, so that $\alpha[\frac{k}{m+1}, \frac{k+1}{m+1}] \subseteq U_k$ maps within one plaque. Let $\varphi_k : \mathbf{R}^{n-q} \times \mathbf{R}^q \to U_k$ be the associated chart. Write $z_k = \alpha(\frac{k}{m+1})$ for $k = 0, \ldots, m+1$ (so $x = z_0$ and $y = z_{n+1}$). Choose transversal sections T_k through z_k such that $T_0 \subseteq U_0$, $T_k \subseteq U_{k-1} \cap U_k$ (for $k = 1, \ldots, m$) and $T_{m+1} \subseteq U_{m+1}$.



These sections T_k are small q-dimensional disks, transverse to the leaf L (and hence transverse to leaves close to L). For instance, if $x = \varphi_0(s,t)$ then one can take for T_0 the image of $\{s\} \times \mathbf{R}^q$ under φ_0. We will now associate to the path α a map germ $(T_0, x) \to (T_{n+1}, y)$, as follows. A point $w_0 \in T_0$ lies on a plaque $P_{w_0} \subseteq U_0$, and this plaque hits T_1 in a unique point w_1, provided w_0 is close enough to x. Next, w_1 lies on a plaque $P_{w_1} \subseteq U_1$ and this plaque hits T_2 in a unique point w_2 (provided w_1 is close enough to z_1, which will be the case if w_0 is close enough to z_0). Proceeding in this way, we find for w_0 close enough to x a sequence $w_k \in T_k$ $(k = 1, \ldots, m+1)$, and we define $\mathrm{hol}(\alpha)(w_0) = w_{n+1} \in T_{n+1}$ as the last point in this sequence. This defined a map germ

$$\mathrm{hol}(\alpha) : (T_0, x) \to (T_{m+1}, y).$$

It is not difficult to prove that $\mathrm{hol}(\alpha)$ does not depend on the choice of the chain $(U_0, \ldots, U_{m+1})$ or on the choice of the intermediate sections T_k $(k = 1, \ldots, m)$. Now define two paths α and β from x to y to be equivalent if $\mathrm{hol}(\alpha) = \mathrm{hol}(\beta)$. The equivalence classes are the arrows from x to y in the holonomy groupoid $\mathrm{Hol}(M, \mathcal{F})$.

If α and β are homotopic (inside the leaf of x and y), then their holonomies are the same. Thus $\mathrm{Hol}(M, \mathcal{F})$ is a quotient of $\mathrm{Mon}(M, \mathcal{F})$, by a homomorphism of groupoids

$$\mathrm{Mon}(M, \mathcal{F}) \to \mathrm{Hol}(M, \mathcal{F}).$$

This holonomy groupoid is a differentiable groupoid; in particular, the domain and codomain maps $\mathrm{Hol}(M, \mathcal{F}) \rightrightarrows M$ are submersions. [See Section 5 for a different construction of the holonomy group, which makes the differentiable structure immediate.]

Exercise. Describe the holonomy and monodromy groupoids in some of the examples given in §1.3 . In particular, observe that in the case of the Möbius band these two groupoids are different (look at the arrows $x \to x$ for a point x on the central leaf).

2.12 Equivalence of topological groupoids. One can adapt the notion of equivalence of categories (in particular, of groupoids) of 2.4 to the topological (or differentiable) context. We will say that a homomorphism $F : C \to D$ between topological groupoids is a *weak equivalence* if

(i) F is essentially surjective, in the sense that

$$d_1 \circ \pi_2 : D_1 \times_{D_0} C_0 \to D_0$$

is an open surjection (where $D_1 \times_{D_0} C_0 = \{(\alpha, x) \mid \alpha \in D_1, x \in C_0 \text{ and } d_0\alpha = F(x)\}$); (ii) F is full and faithful, in the sense that the square

$$\begin{array}{ccc} C_1 & \xrightarrow{F} & D_1 \\ {\scriptstyle (d_0,d_1)}\downarrow & & \downarrow{\scriptstyle (d_0,d_1)} \\ C_0 \times C_0 & \xrightarrow{F\times F} & D_0 \times D_0 \end{array}$$

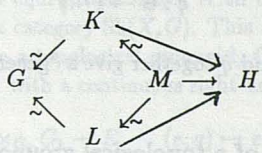
is a pullback (fibered product).

The definition for differentiable groupoids is similar, except that one replaces open surjection in (i) by *surjective submersion.*

2.13 Inverting weak equivalences. One can construct a category of topological groupoids and "generalized" homomorphisms, by formally inverting the weak equivalences. Such a generalized homomorphism $G \to H$ is represented by two (ordinary) homomorphisms

$$G \xleftarrow{\sim} K \to H$$

where the first is a weak equivalence. Two such diagrams $G \xleftarrow{\sim} K \to H$ and $G \xleftarrow{\sim} L \to H$ are said to be equivalent (i.e. represent the same generalized homomorphism $G \to H$) when there exists a diagram of topological groupoids



where the homomorphisms from M into K and L are weak equivalences as indicated, and the diagram commutes up to continuous natural isomorphisms. Such generalized homomorphisms can be composed by a construction just like that for a "category of fractions" [Gabriel-Zisman (1967)]. Of course a similar construction applies to the category of differentiable groupoids. We will come back to this construction in 3.14 and 5.4.

2.14 An etale version of holonomy. An important example of a weak equivalence between topological (differentiable) groupoids is the following. Let $(M, \mathcal{F})$ be a foliation of codimension q, and let $T \hookrightarrow M$ be a complete transversal. This means that T is an immersed q-dimensional submanifold of M which is transversal to the leaves of the foliation $\mathcal{F}$, and intersects each leaf at least once. (For example, T could be the disjoint union of a sufficiently large collection of small transversal sections as considered in 2.11.) Define a new groupoid $\mathrm{Hol}_T(M, \mathcal{F})$ which has T as space of objects, and exactly the same arrows as $\mathrm{Hol}(M, \mathcal{F})$. In other words, the space $\mathrm{Hol}_T(M, \mathcal{F})_1$ of arrows of $\mathrm{Hol}_T(M, \mathcal{F})$ fits by definition into a pullback

$$\begin{array}{ccc} \mathrm{Hol}_T(M,\mathcal{F})_1 & \hookrightarrow & \mathrm{Hol}(M,\mathcal{F})_1 \\ {\scriptstyle (d_0,d_1)}\downarrow & & \downarrow{\scriptstyle (d_0,d_1)} \\ T \times T & \hookrightarrow & M \times M. \end{array}$$

The immersion $T \hookrightarrow M$ gives a homomorphism of topological groupoids

$$\mathrm{Hol}_T(M,\mathcal{F}) \to \mathrm{Hol}(M,\mathcal{F}),$$

and this homomorphism is a weak equivalence. The groupoid $\mathrm{Hol}_T(M,\mathcal{F})$ has the special property that the domain and codomain maps are etale: it is an etale groupoid (cf. 2.7).

2.15 Mapping the holonomy into Γ^q. For another example of a weak equivalence, recall from 1.4.1 that the foliation $\mathcal{F}$ on M can be represented by submersions $\psi_i : U_i \to \mathbf{R}^q$ for an open cover $M = \bigcup U_i$. Let $G_0 = \sum U_i$ be the disjoint union of all the U_i, and let $\pi : G_0 \twoheadrightarrow M$ be the local diffeomorphism given by the inclusions $U_i \hookrightarrow M$. Define G_1 as the pullback

$$\begin{array}{ccc} G_1 & \longrightarrow & \mathrm{Hol}(M,\mathcal{F})_1 \\ \downarrow & & \downarrow \\ G_0 \times G_0 & \xrightarrow{\pi\times\pi} & M \times M. \end{array}$$

Then G_1 is the space of the arrows of a differentiable groupoid G, where an arrow $x \to y$ in G is by definition given by a unique arrow $\pi x \to \pi y$ in $\mathrm{Hol}(M,\mathcal{F})$. The map π gives rise to a weak equivalence

$$\pi : G \xrightarrow{\sim} \mathrm{Hol}(M,\mathcal{F}). \tag{2.15.1}$$

Note also that the maps $\pi_1 : U_i \to \mathbf{R}^q$ and the maps h_{ij} occurring in 1.4.1 give rise to a homomorphism of groupoids

$$q : G \to \Gamma^q\,. \tag{2.15.2}$$

These two homomorphisms π and q together give a "generalized" homomorphism $\mathrm{Hol}(M,\mathcal{F}) \to \Gamma^q$, as in 2.13.

2.16 The classifying space of a topological groupoid. (This also can be done for any topological category.) For a topological groupoid G one can construct a topological space BG, called the classifying space of G. This construction is discussed in detail e.g. in [Segal (1968), Bott (1972)]. Briefly, one first constructs a simplicial space, i.e. a functor

$$\mathrm{Nerve}(G) : \mathbb{\Delta}^{\mathrm{op}} \to (\text{spaces}).$$

One writes $\mathrm{Nerve}(G)_n$ for $\mathrm{Nerve}(G)([n])$. For $n = 0,1$, define $\mathrm{Nerve}(G)_0 = G_0$, and $\mathrm{Nerve}(G)_1 = G_1$. In general, $\mathrm{Nerve}(G)_n$ is defined as the space of composable strings of arrows

$$x_0 \xleftarrow{\alpha_1} x_1 \leftarrow \ldots \xleftarrow{\alpha_n} x_n$$

in G (topologized as the fibered product $G_1 \times_{G_0} \cdots {}_{G_0} \times G_1$). If $\varphi : [n] \to [m]$ is an order preserving map (an arrow in $\mathbb{\Delta}$), then the functor $\mathrm{Nerve}(G)$ takes φ to a map denoted

$$\varphi^* : \mathrm{Nerve}(G)_m \to \mathrm{Nerve}(G)_n$$

and defined from the composition and identities of G as follows: the value of φ^* for a string $(x_0 \xleftarrow{\alpha_1} \ldots \xleftarrow{\alpha_n} x_n)$ is the string $(x_{\varphi(0)} \leftarrow x_{\varphi(1)} \leftarrow \ldots \leftarrow x_{\varphi(n)})$, where the arrow $x_{\varphi(i+1)} \to x_{\varphi(i)}$ is either the identity arrow (if $\varphi(i+1) = \varphi(i)$) or the composition $\alpha_{\varphi(i+1)} \circ \ldots \circ \alpha_{\varphi(i)+1}$ (if $\varphi(i) < \varphi(i+1)$).

The space BG is defined as the geometric realization of this simplicial space $\mathrm{Nerve}(G)$.

This construction is functorial: a homomorphism $G \to H$ if topological groupoids gives continuous map $BG \to BH$ between their classifying spaces. Moreover, we quote from Haefliger (1984):

2.16.1 Proposition. *If $G \to H$ is a weak equivalence of topological groupoids then the induced map $BG \to BH$ is a weak homotopy equivalence.*

3 Grotendieck étendues.

3.1 Examples of topoi. (i) For a group G, the category of all (right) G-sets, and equivariant (i.e., action-preserving) functions between them, is a topos, denoted G-**Sets**.

(ii) For a small category C (one for which C_0 and C_1 are sets), a presheaf on C is a functor $P : C^{\mathrm{op}} \to$ **Sets**. These presheaves form a category $\hat{C}$, with the natural transformations as arrows. This category $\hat{C}$ is a topos. (Note that when we view a group as a category with one object, then example (i) is a special case of example (ii), since (G-**Sets**) $= \hat{G}$.)

(iii) Let X be a topological space. A *sheaf* on X is a local homeomorphism (etale map) $p : E \to X$. An arrow between such sheaves is a continuous function f over X:

$$\begin{array}{ccc} E & \xrightarrow{f} & E' \\ & {\scriptstyle p}\searrow \quad \swarrow{\scriptstyle p'} & \\ & X & \end{array} \qquad p' \circ f = p\,.$$

This gives a category $\mathrm{Sh}(X)$ of all sheaves on X, which is a topos.

(iv) Generalizing examples (i) and (iii), let G be a group and let X be a space equipped with a continuous right G-action $X \times G \to X$, denoted $(x, g) \mapsto x \cdot g$. A G-sheaf (or *equivariant sheaf*) on X is a sheaf $p : E \to X$, equipped with a continuous G-action $E \times G \to E$ making p into an equivariant map. With equivariant continuous functions over X, these G-sheaves form a category $\mathrm{Sh}(X, G)$. This category is a topos.

(v) More generally, let G be a topological groupoid, $G = (G_1 \rightrightarrows G_0$, etc.). A G-sheaf is a sheaf $p : E \to G_0$ equipped with a continuous right action by G, denoted

$$E \times_{G_0} G_1 \to E \qquad (e, g) \mapsto e \cdot g\,.$$

Thus $e \cdot g$ is defined whenever $p(e) = d_1(g)$ and satisfies the identity $p(e \cdot g) = d_0(g)$ as well as the usual identities for an action

$$(e \cdot g) \cdot h = e \cdot (g \cdot h) \quad \text{and} \quad e \cdot i(x) = e\,,$$

for any $e \in E$, $g, h \in G_1$ and $x \in G_0$ for which these expressions are defined. A G-map between two such G-sheaves E and F is a continuous map which respects the etale projections into X as well as the action, thus making the diagrams

$$\begin{array}{ccc} E & \longrightarrow & F \\ & {\scriptstyle p}\searrow \quad \swarrow{\scriptstyle p'} & \\ & X & \end{array} \qquad \begin{array}{ccc} E \times_{G_0} G_1 & \xrightarrow{f \times \mathrm{id}} & F \times_{G_0} G_1 \\ \downarrow & & \downarrow \\ E & \xrightarrow{f} & F \end{array}$$

commute. In this way one obtains a category $\mathcal{B}G$ of all G-sheaves. This category is a topos, called the *classifying topos* of G.

Example (iv) is a special case: if G is a group acting on a space X, one can form the *"translation groupoid"* X_G with X as space of objects and $X \times G$ as space of arrows: an arrow $x \to y$ is a pair (y, g) where $y \cdot g = x$. Then $\mathcal{B}(X_G) = \mathrm{Sh}(X, G)$.

The construction of $\mathcal{B}G$ from G is functorial in G, as will be discussed in 3.13 and 3.14 below. Here we note the following:

3.2 Proposition. *A weak equivalence $G \to H$ between topological groupoids induces an equivalence of categories $\mathcal{B}G \cong \mathcal{B}H$.*

3.3 Definition of topos. There are various equivalent definitions of a (Gothendieck) topos. One is as the category of all sheaves on a site ([SGA4], Mac Lane-Moerdijk, p. 127]). Another one is as a category satisfying the Giraud axioms [Mac Lane-Moerdijk, p. 575]. A third one is as a category which has a set of generators as well as all sums, and satisfies the Lawvere-Tierney axioms [Mac Lane-Moerdijk, p. 591]. The precise form of the definition does not matter much here. What is important is, first, the list of examples in 3.1. Secondly,that one should *not* think of a topos as a very large category, *but* as a generalized kind of space. For example, the topos $\mathrm{Sh}(X)$ of 3.1(iii)"is" really the space X in disguise. This becomes clear when we define maps between topoi (in 3.12 below), in such a way that topos maps $\mathrm{Sh}(X) \to \mathrm{Sh}(Y)$ correspond to continuous functions $X \to Y$. This intuition is further supported by the homotopy and cohomology groups of a topos (§4) which generalize those of a space. Thirdly, for the definition of a map between topoi it is important that a topos is a category with (among other things) arbitrary (small) colimits and finite limits. Finally, we may remark that Joyal and Tierney (1984) prove a theorem which almost expresses that *every* topos is of the form $\mathcal{B}G$ as in 3.1(v) [the only difference being that one must consider not just topological groupoids but also localic groupoids; see also Moerdijk (1988a),(1988b), Joyal-Moerdijk(1990).]

3.4 Slice topoi. If $\mathcal{C}$ is any category and E is an object of $\mathcal{C}$ then the category $\mathcal{C}/E$ is defined as having all arrows $p : C \to E$ in $\mathcal{C}$ as objects, and having as arrows all commutative triangles

$$\begin{array}{ccc} C & \longrightarrow & C' \\ & {\scriptstyle p}\searrow \quad \swarrow{\scriptstyle p'} & \\ & E & \end{array}$$

If $\mathcal{C}$ is a topos, then so is $\mathcal{C}/E$. For example, if $E \to X$ is a sheaf on a space X, then (up to equivalence of categories) $\mathrm{Sh}(X)/E$ is the category $\mathrm{Sh}(E)$ of all sheaves on E. And if S is a G-set, then the category $(G\text{-}\mathbf{Sets})/S$ is the topos $\mathcal{B}(S_G)$ (see 3.1(v)), where S_G is the discrete groupoid with S as set of objects, and as arrows $s \to s'$ all those $g \in G$ with the property that $s' \cdot g = s$. Since S_G is discrete, we may also write $\mathcal{B}(S_G)$ as the presheaf topos $\hat{S}_G$ (cf. 3.1(ii)).

3.5 Definition of étendue. (SGA4) An *étendue* is a topos $\mathcal{T}$ with the property that there exists an object $E \in \mathcal{T}$ such that the unique arrow $E \to 1_{\mathcal{T}}$ into the terminal object $1_{\mathcal{T}}$ of $\mathcal{T}$ is an epimorphism, and such that $\mathcal{T}/E$ is equivalent to the topos $\mathrm{Sh}(X)$ of sheaves on some space X. Thus an étendue is a topos which is "locally" like a space. For more concrete descriptions of étendues, see Theorems 3.7 and 3.10 below.

3.6 Examples. The easiest examples of étendues are the following two. Other examples are provided by Theorems 3.7 and 3.10.

(i) For any space X, the topos $\mathrm{Sh}(X)$ is an étendue (take E to be the terminal sheaf $\mathrm{id} : X \to X$ itself).

(ii) For any group G, the topos $(G\text{-}\mathbf{Sets})$ is an étendue: Let $\tilde{G}$ be the group G viewed as a right G-set. Then the translation groupoid $(\tilde{G})_G$ is a groupoid with exactly one arrow between any two objects. Consequently the slice topos $(G\text{-}\mathbf{Sets})/\tilde{G}$, which is $\mathcal{B}(\tilde{G}_G)$ by 3.4, is equivalent to the category **Sets**. But **Sets** is sheaves on the one-point space, so $(G\text{-}\mathbf{Sets})$ is an étendue.

3.7 Theorem. (Grothendieck-Verdier) *A topos $\mathcal{T}$ is an étendue iff $\mathcal{T}$ is (equivalent to) the classifying topos $\mathcal{B}G$ of an etale topological groupoid.*

(Recall from §2 that a groupoid is said to be etale if its domain and codomain maps are etale.)

3.8 Examples. (i) The Haefliger groupoid Γ^q (see 2.6) is etale, so its classifying topos $\mathcal{B}\Gamma^q$ is an étendue.

(ii) The holonomy groupoid $\mathrm{Hol}(M,\mathcal{F})$ of a foliated manifold is weakly equivalent to an etale groupoid (namely $\mathrm{Hol}_T(M,\mathcal{F})$, cf. 2.14). By 3.2 and 3.7 it follows that its classifying topos $\mathcal{B}\,\mathrm{Hol}(M,\mathcal{F})$ is an étendue.

(iii) For a space X equipped with a right G-action, the topos $\mathrm{Sh}(X,G)$ of equivariant sheaves, considered in 3.1(iv), is an étendue. Indeed, $\mathrm{Sh}(X,G) = \mathcal{B}(X_G)$ as remarked at the end of 3.1, and the translation groupoid X_G is etale.

3.9 Local equivalence relations. Recall from §1.4 that a local equivalence relation $\underline{r}$ on a space M is given by an "atlas" $\{(U_i, R_i)\}$ where $M = \bigcup U_i$ is an open cover of M and each R_i is an equivalence relation on U_i. We call $\underline{r}$ *locally simply connected* if there exists such an atlas $\{(U_i, R_i)\}$ for $\underline{r}$ with the following properties, for each index i:

(i) the two projections $R_i \rightrightarrows U_i$ are open maps;

(ii) for each $x \in U_i$ its equivalence class xR_i is a simply connected subset of U_i.

For example, the local equivalence relation on M defined by a foliation of codimension q is evidently locally simply connected, since inside a chart U_i the equivalence classes are the plaques, all diffeomorphic to $\mathbf{R}^{n-q}$.

For such a local equivalence relation $\underline{r}$, a sheaf $p : E \to M$ on M is said to be $\underline{r}$-*invariant* if for each chart (U_i, R_i) with simply connected equivalence classes as above, the sheaf E restricts to a constant sheaf (= a trivial covering projection) on each equivalence class $xR_i \subseteq U_i$. For example, if $\underline{r}$ is defined by a foliation on M then a sheaf E is $\underline{r}$-invariant iff for every leaf $L \subseteq M$ the restriction $p^{-1}(L) \to L$ is a covering projection.

These invariant sheaves on M form a full subcategory of $\mathrm{Sh}(M)$,denoted

$$\mathrm{Sh}(M,\underline{r})\,.$$

This category of $\underline{r}$-invariant sheaves is a topos. In fact it is an étendue, and every étendue is of this form:

3.10 Theorem. (Kock-Moerdijk) *A topos $\mathcal{T}$ is an étendue iff $\mathcal{T}$ is (equivalent to) a category of the form* $\mathrm{Sh}(M,\underline{r})$, *for some locally simply connected local equivalence relation $\underline{r}$ on a space M.*

3.11 Monodromy of locally equivalence relations. Recall from 2.10 the monodromy groupoid $\mathrm{Mon}(M,\mathcal{T})$ of a foliation. More generally, one can define in exactly the same way a topological groupoid

$$\mathrm{Mon}(M,\underline{r})$$

for any (locally simply connected) local equivalence relation $\underline{r}$ on a space M. Its space of objects is again M. An arrow $x \to y$ in $\mathrm{Mon}(M,\underline{r})$ is a homotopy class of paths $\alpha : [0,1] \to M$ from x to y, which locally maps within one equivalence class [i.e. there is a chain

$(U_0, R_0), \ldots, (U_m, R_m)$ of charts for $\underline{r}$ so that $\alpha[\frac{k}{m+1}, \frac{k+1}{m+1}] \subseteq U_k$ and $<\alpha(t), \alpha(t')> \in R_k$ for $\frac{k}{m+1} \leq t,\ t' \leq \frac{k+1}{m+1}$.] Furthermore, the homotopies between such paths, used in the definition of the arrows in $\mathrm{Mon}(M, \underline{r})$, must similarly lie locally within one equivalence class. In Kock-Moerdijk (1991) the following result is proved:

Proposition. *There is an equivalence of topoi*

$$\mathrm{Sh}(M, \underline{r}) \cong \mathcal{B}\,\mathrm{Mon}(M, \underline{r})\,.$$

In other words, for a suitable groupoid the topos $\mathrm{Sh}(M, \underline{r})$ is one of our typical examples 3.1(v).

3.12 Mappings between topoi. For two topoi $\mathcal{T}$ and $\mathcal{T}'$, a map $f : \mathcal{T} \to \mathcal{T}'$ can be defined in (at least) two equivalent ways:

(a) as a pair of functors $f_* : \mathcal{T} \to \mathcal{T}'$ and $f^* : \mathcal{T}' \to \mathcal{T}$ such that f^* preserves finite limits and f_* is right adjoint to f^*,

(b) as a functor $f^* : \mathcal{T}' \to \mathcal{T}$ which preserves colimits and finite limits.

The functor f^* is called the *invers image* of the map f, and f_* is called the *direct image*. Two such maps f and $g : \mathcal{T} \to \mathcal{T}'$ are said to be isomorphic if there exists a natural isomorphism (cf. 2.3) $\tau : f^* \xrightarrow{\sim} g^*$ between their inverse image functors. Isomorphic maps are often identified in practice. We write $\mathrm{Hom}(\mathcal{T}, \mathcal{T}')$ for the collection of *isomorphism classes* of maps from $\mathcal{T}$ to $\mathcal{T}'$.

3.13 Examples. (i) Let X, Y be topological spaces. A continuous map $f : X \to Y$ induces a mapping of topoi, (again denoted) $f : \mathrm{Sh}(X) \to \mathrm{Sh}(Y)$, with the usual inverse and direct image functors

$$f^* : \mathrm{Sh}(Y) \rightleftarrows \mathrm{Sh}(X) : f_*$$

(see any book on sheaf theory). If Y is Hausdorff then any map of topoi $\mathrm{Sh}(X) \to \mathrm{Sh}(Y)$ is induced in this way by a unique continuous map $X \to Y$.

(ii) Let G and H be groups. A homomorphism of groups $\varphi : G \to H$ induces a map of topoi $\varphi : (G\text{-}\mathbf{Sets}) \to (H\text{-}\mathbf{Sets})$. The inverse image functor φ^* sends an H-set S to the same set S viewed as a G-set (with action by G defined in terms of the given action by H as $s \cdot g := s \cdot \varphi(g)$). The direct image functor φ_* sends a G-set R to the set

$$\mathrm{Hom}_G(H, R)$$

of G-equivariant maps (where the G-action on H is defined as $h \cdot g := h \cdot \varphi(g)$). The group H acts on elements α of this set $\mathrm{Hom}_G(H, R)$ as

$$(\alpha \cdot h)(k) = \alpha(hk)\,.$$

It is not difficult to prove that any mapping of topoi $(G\text{-}\mathbf{Sets}) \to (H\text{-}\mathbf{Sets})$ is induced in this way (up to isomorphism) from a homomorphism of groups $G \to H$ (unique up to conjugation).

(iii) Let G and H be topological groupoids and let $\varphi : G \to H$ be a homomorphism (i.e., a continuous functor). Such a homomorphism induces a map of classifying topoi, denoted

$$\varphi : \mathcal{B}G \to \mathcal{B}H\,.$$

We describe the inverse image functor $\varphi^* : \mathcal{B}H \to \mathcal{B}G$ explicitly (in other words, we use version (b) of the definition 3.12). The direct image functor φ^* is harder to describe explicitly. Let F be an H-sheaf, given by an etale projection $p : F \to H_0$ and an action $F \times_{H_0} H_1 \to F$, as in 3.1(v). Construct the fibered product

$$\begin{array}{ccc} G_0 \times_{H_0} F & \longrightarrow & F \\ {\scriptstyle \pi_1}\downarrow & & \downarrow{\scriptstyle p} \\ G_0 & \xrightarrow{\varphi} & H_0 \end{array}$$

Thus $G_0 \times_{H_0} F$ consits of all pairs (x, y) where $x \in G_0$ and $y \in F$ are such that $\varphi(x) = p(y)$. The projection $\pi_1 : G_0 \times_{H_0} F \to G_0$ is again etale, since p is. Moreover, there is an action of the groupoid G on this pullback $G_0 \times_{H_0} F$, defined for any arrow $g : x' \to x$ in G and any point $(x, y) \in G_0 \times_{H_0} F$ by using the action of H on F as

$$(x, y) \cdot g = (x', y \cdot \varphi(g)) \, .$$

In this way, the space $G_0 \times_{H_0} F$ with its projection π_1 and G-action defines a G-sheaf, denoted $\varphi^*(F)$. This yields a functor $\varphi^* : \mathcal{B}H \to \mathcal{B}G$.

Note that this construction generalizes the case of groups. However, unlike the case of groups, it is not generally true for topological groupoids that a map of classifying topoi $\mathcal{B}G \to \mathcal{B}H$ is induced by a homomorphism of groupoids $G \to H$. Nonetheless, this is *almost* true if G and H are etale topological groupoids, as we will now discuss.

3.14 The localization theorem for étendues. *Let G and H be etale topological groupoids (or groupoids which are weakly equivalent to such, 2.12). Then there is a bijective correspondence between mappings of classifying topoi $\mathcal{B}G \to \mathcal{B}H$ and equivalence classes of generalized homomorphisms of groupoids as described in* 2.13.

Using the Grothendieck-Verdier theorem 3.7 it follows that the category of étendues is precisely the one obtained from the category of etale topological groupoids by inverting the weak equivalences:

$$(\text{étendues}) \quad \cong \quad (\text{etale groupoids})[w.e.^{-1}].$$

The category on the right can be constructed as a category of fractions in the sense of [Gabriel-Zisman]. This result 3.14 is a special case of the localization theorem for general topoi proved in [Moerdijk (1988a)]; see also [Moerdijk (1988b), Pronk (in preparation)].

We will come back to this localization theorem in §5.4.

4 Homotopy and cohomology of étendues.

4.1 Cohomology of topoi. [SGA4] For any topos $\mathcal{T}$ and any abelian group object A of $\mathcal{T}$, there are cohomology groups

$$H^n(\mathcal{T}, A) \qquad (n \geq 0).$$

The construction is functorial in $\mathcal{T}$ and A: a homomorphism $A \to B$ induces homomorphisms $H^n(\mathcal{T}, A) \to H^n(\mathcal{T}, B)$, while a mapping of topoi $f : \mathcal{T}' \to \mathcal{T}$ induces homomorphisms $f^* : H^n(\mathcal{T}, A) \to H^n(\mathcal{T}', f^*(A))$. In particular, if f is an equivalence of topoi then f^* is an isomorphism $H^n(\mathcal{T}, A) \cong H^n(\mathcal{T}', f^*(A))$. [One way to define these groups is via the global sections functor $\Gamma : \mathcal{T} \to \mathbf{Sets}$, which is defined as $\Gamma(E) = \mathrm{Hom}_{\mathcal{T}}(1_{\mathcal{T}}, E)$, the set of all arrows in $\mathcal{T}$ from the terminal object $1_{\mathcal{T}}$ to E. Let $\mathrm{Ab}(\mathcal{T})$ be the category of abelian group objects in $\mathcal{T}$. This is an abelian category with enough injectives. The functor Γ gives a functor $\Gamma : \mathrm{Ab}(\mathcal{T}) \to \mathrm{Ab}(\mathbf{Sets})$, and the cohomology groups of $\mathcal{T}$ are defined as the right derived functors of the latter functor Γ, viz. $H^n(\mathcal{T}, A) = R^n\Gamma(A)$.]

4.2 Examples. (i) For a space X, an abelian group object of the topos $\mathrm{Sh}(X)$ is a sheaf of abelian groups on X. (This is a sheaf $p : A \to X$ such that each fiber $p^{-1}(x) = A_x$ has the structure of an abelian group, and such that these structures are compatible in the sense that the induced maps $A \times_X A \xrightarrow{+} A$ and $X \xrightarrow{0} A$ are continuous.) For such a sheaf A, the topos cohomology groups $H^n(\mathrm{Sh}(X), A)$ are the ordinary sheaf cohomology groups $H^n(X, A)$, studied e.g. in [Godement (1958)]. If B is an abelian group (in **Sets**), there is an associated *constant* abelian sheaf $\Delta(B) = (X \times B \xrightarrow{\pi_1} X)$ on X, and the sheaf cohomology groups $H^n(X, \Delta B)$ are isomorphic to the usual singular cohomology groups $H^n_{\mathrm{sing}}(X, B)$, provided X is a "good" space (e.g. a manifold).

(ii) For a group G, an abelian group object in the topos (G-**Sets**) (see 3.1(i)) is an abelian group A with an action by G (or: a $\mathbf{Z}[G]$-module). For such an abelian group A, the topos-cohomology groups $H^n(G\text{-}\mathbf{Sets}, A)$ are the Eilenberg-MacLane cohomology groups $H^n(G, A)$ of the group G, studied e.g. in [Brown (1982)].

4.3 Cohomology of étendues. Let $\mathcal{T}$ be an étendue. By the Grothendieck-Verdier theorem 3.7, there exists an etale topological groupoid G so that $\mathcal{T}$ is (equivalent to) the topos $\mathcal{B}G$ of G-sheaves. (And, by the localization theorem 3.14, this groupoid G is unique up to weak equivalence.) An abelian group object A in $\mathcal{B}G$ is a sheaf of abelian groups $p : A \to G_0$ over G_0 on which G acts by group homomorphisms; in other words, the action map $A \times_{G_0} G_1 \to A$ has the property that, for any arrow $g : x \to y$ in G, the action by g gives a group homomorphism (in fact, an isomorphism) $A_y \to A_x$. For such an A, the cohomology groups $H^n(\mathcal{B}G, A)$ are related to the sheaf cohomology groups of the spaces $\mathrm{Nerve}_p(G)$ (for $p \geq 0$) described in 4.2(i), via a spectral sequence

$$H^p H^q(\mathrm{Nerve}_p(G), A^{(p)}) \Rightarrow H^{p+q}(\mathcal{B}G, A)\,. \tag{4.3.1}$$

Here $A^{(p)}$ denotes the sheaf on $\mathrm{Nerve}_p(G)$ induced by pullback along the canonical map of topoi $\mathrm{Sh}(\mathrm{Nerve}_p(G)) \to \mathcal{B}G$; explicitly, for a point $g = (x_0 \xleftarrow{g_1} x_1 \leftarrow \ldots \xleftarrow{g_p} x_p)$ of $\mathrm{Nerve}_p(G)$, the fiber of $A^{(p)}$ over g is A_{x_0}.

4.4 Locally connected topoi. Let $\mathcal{T}$ be a topos. An object E of $\mathcal{T}$ is *connected* if $E \neq 0$ (= the initial object of $\mathcal{T}$) and if E cannot be decomposed as a sum $E \cong E_1 + E_2$, except in the trivial ways $E \cong E + 0$ and $E \cong 0 + E$. The topos $\mathcal{T}$ is said to be *connected* if the terminal object $1_{\mathcal{T}}$ is a connected object. The topos $\mathcal{T}$ is called *locally connected* if every object E can be decomposed as a sum of connected objects, say $E \cong \sum_{i \in I} E_i$. This decomposition is essentially unique, and its index-set I is the set of *connected components* of E, denoted $\pi_0(E)$. In this way, one obtains a functor

$$\pi_0 : \mathcal{T} \to \mathbf{Sets}\ .$$

4.5 Examples. (i) For a locally connected space X, the topos of sheaves $\mathrm{Sh}(X)$ is a locally connected topos. Indeed, if $p : E \to X$ is a sheaf then E is again a locally connected space, and $\pi_0(E)$ is its set of connected components, in the usual topological sense.

(ii) Let G be a group. The topos (G-**Sets**) is locally connected. For a G-set S, its connected components are the orbits $\{s \cdot g \mid g \in G\}$, for any $s \in S$; thus, $\pi_0(S) = S/G$.

(iii) Let G be an etale topological groupoid, with associated étendue $\mathcal{B}G$ of G-sheaves. If the space G_0 of objects is locally connected, then so is the space G_1 of arrows, since $d_0 : G_1 \to G_0$ is assumed to be a local homeomorphism. In this case the topos $\mathcal{B}G$ is locally connected. For a sheaf $p : E \to G_0$ with G-action $E \times_{G_0} G_1 \to E$, the set of connected components $\pi_0(E)$ is obtained from the ordinary topological set of connected components $\pi_0^{\mathrm{top}}(E)$ of the space E, by identifying the component of e and that of $e \cdot g$, for any point $e \in E$ and any arrow $g : x \to y$ in G (with $y = p(e)$ so that $e \cdot g$ is defined). One may write, suggestively, $\pi_0(E) = \pi_0^{\mathrm{top}}(E)/G$.

4.6 The fundamental group. (sketch) Let $\mathcal{T}$ be a topos. An object E of $\mathcal{T}$ is called *locally constant* if there exists an object $U \in \mathcal{T}$ with the property that the unique map $U \to 1_{\mathcal{T}}$ into the terminal object $1_{\mathcal{T}}$ of $\mathcal{T}$ is an epimorphism, and such that $E \times U$ is isomorphic to a sum of copies of U,

$$E \times U \cong \sum_{i \in I} U \qquad \text{(for some index set } I\text{)},$$

by an isomorphism *over* U (i.e., one which makes the diagram

$$\begin{array}{ccc} E \times U & \xrightarrow{\sim} & \sum_i U \\ & {\scriptstyle \pi_2}\searrow \quad \swarrow{\scriptstyle \varphi} & \\ & U & \end{array}$$

commute, where φ is the identity on each summand). The object E is called *constant* when one can take $U = 1_{\mathcal{T}}$.

Write $\Pi_1(\mathcal{T})$ for the full subcategory of $\mathcal{T}$ consisting of sums of such locally constant objects. If $\mathcal{T}$ is a connected and locally connected topos, it can be shown that there is a unique (up to isomorphism) pro-group G such that $\Pi_1(\mathcal{T})$ is equivalent to the category (G-**Sets**) of sets with a continuous G-action. This unique G is by definition the *fundamental group* of the topos $\mathcal{T}$, and denoted $\pi_1(\mathcal{T})$. In our examples below, $\mathcal{T}$ wil generally be locally *simply* connected and $\pi_1(\mathcal{T})$ will be an ordinary group (not a pro-group). [Note: exactly as for topological spaces, one should really define $\pi_1(\mathcal{T}, p)$ where p is a "base-point" of $\mathcal{T}$, i.e. a map of topoi p : **Sets** $\to \mathcal{T}$. If $\mathcal{T}$ is connected, then $\pi_1(\mathcal{T}, p)$ and $(\pi_1(\mathcal{T}, p')$ are isomorphic for any two such base-ponts p and p'; however, this isomorphism is not canonical.]

For more on fundamental groups of topoi, see e.g. Grothendieck (SGA1), Barr and Diaconescu (1981), Kennison (1990), Moerdijk(1989).

4.7 Examples. (i) Let X be a locally connected space. A sheaf $p : E \to X$ is a locally constant object of the topos $\mathrm{Sh}(X)$ iff p is a covering projection. If X is connected and locally simply connected, then for any base-point $x_0 \in X$ there is a canonical isomorphism $\pi_1(\mathrm{Sh}(X), x_0) \cong \pi_1(X, x_0)$ (where on the left x_0 is viewed as a point of the topos $\mathrm{Sh}(X)$). In other words, the toposophic definition of the fundamental group agrees with the usual one.

(ii) Let G be a group. Any G-set S is a locally constant object of the topos G-**Sets** (take $U = \tilde{G}$ in the definition 4.6, where $\tilde{G}$ is G as a right G-set, cf. 3.6(ii)). Thus G is the fundamental group of the topos G-**Sets**.

(iii) Let G be an etale topological groupoid, and assume G_0 locally connected, as in 4.3(iii). A G-sheaf E is locally constant as an object of $\mathcal{B}G$ iff $p : E \to G_0$ is a covering projection (cf. [Moerdijk (1991), Lemma 4.2]). This observation allows us to give an explicit description of the fundamental group of $\mathcal{B}G$ in case G_0 is locally simply connected (e.g. if G_0 is a manifold). The details are given in [Moerdijk(1991), §4]. In the case where G is an S-atlas (see 2.8) this description agrees with the one given by Van Est (1984). See also §5.2.

4.8 Hurewicz formula. The fundamental group of a connected and locally connected topos $\mathcal{T}$ shares many of the properties of fundamental groups of spaces. For instance, for any abelian group A there is an isomorphism

$$H^1(\mathcal{T}, A) \cong \mathrm{Hom}(\pi_1(\mathcal{T}), A)$$

(where on the left, A is identified with the associated abelian group object $\Delta(A) = \sum_{a \in A} 1_{\mathcal{T}}$ of $\mathcal{T}$).

4.9 Etale homotopy. For a locally connected topos $\mathcal{T}$ and a base-point p of $\mathcal{T}$ one can also introduce higher homotopy groups $\pi_n(\mathcal{T}, p)$ (all $n \in \mathbb{N}$), called etale homotopy groups. The construction is described in detail in [Artin-Mazur (1969)]. These groups depend functorially on $\mathcal{T}$, in the sense that a map $f : \mathcal{T} \to \mathcal{T}'$ induces group homomorphisms $\pi_n(\mathcal{T}, p) \to \pi_n(\mathcal{T}', f(p))$ for all $n \geq 0$. If $\mathcal{T} = \mathrm{Sh}(X)$ is a topos of sheaves on a locally contractible space (e.g. a manifold), then these groups are isomorphic to the usual homotopy groups of the space X, by the comparison theorem of [Artin-Mazur, §12].

4.10 Weak homotopy type. For locally connected topoi with "enough points" (at least one point in each connected component), a map $f : \mathcal{T} \to \mathcal{T}'$ is said to be a weak homotopy equivalence if for any point p of $\mathcal{T}$, the map f induces an isomorphism $\pi_n(\mathcal{T}, p) \xrightarrow{\sim} \pi_n(\mathcal{T}', f(p))$, for any $n \geq 0$. By the Artin-Mazur comparison theorem, this agrees with the usual notion of weak homotopy equivalence between spaces, for a map between locally contractible spaces $X \to X'$ and the induced map of topoi $\mathrm{Sh}(X) \to \mathrm{Sh}(X')$. Two topoi $\mathcal{T}$ and $\mathcal{T}'$ are said to be of the same weak homotopy type if there exists a zig-zag of weak homotopy equivalences $\mathcal{T} \to \cdot \leftarrow \cdot \to \cdot \leftarrow \ldots \leftarrow \mathcal{T}'$.

4.11 Toposophic Whitehead theorem. (Artin-Mazur) *A map $f : \mathcal{T} \to \mathcal{T}'$ between connected and locally connected topoi is a weak homotopy equivalence iff f induces an isomorphism of fundamental groups as well as isomorphisms $H^n(\mathcal{T}', A) \to H^n(\mathcal{T}, f^*A)$ (all $n \geq 0$) for any locally constant abelian group A in $\mathcal{T}'$.*

4.12 Comparison theorem for étendues. [Moerdijk (1991)] *Let G be an etale topological groupoid, with associated classifying space BG and classifying topos $\mathcal{B}G$. There is a canonical weak homotopy equivalence* $\mathrm{Sh}(BG) \xrightarrow{\sim} BG$.

(Recall that if BG is a locally contractible space then the (etale) homotopy groups and the cohomology groups with locally constant coefficients of the topos $\mathrm{Sh}(BG)$ are identical to those of the space BG. Thus the theorem expresses that the space BG has the same weak homotopy type as the topos $\mathcal{B}G$.)

4.13 Homotopy of maps. Let X be a space and $\mathcal{T}$ a topos. Two topos-maps $f, g : \mathrm{Sh}(X) \rightrightarrows \mathcal{T}$ are said to be homotopic if there exists a topos map $h : \mathrm{Sh}(X \times [0,1]) \to \mathcal{T}$ such that for the inclusions $i_k : X \hookrightarrow X \times [0,1]$ $(k = 0, 1)$ and their associated topos maps $i_k : \mathrm{Sh}(X) \hookrightarrow \mathrm{Sh}(X \times [0,1])$ there are isomorphisms $h \circ i_0 \cong f$ and $h \circ i_0 \cong g$. The set of homotopy classes of maps $\mathrm{Sh}(X) \to \mathcal{T}$ is denoted by $[\mathrm{Sh}(X), \mathcal{T}]$. (One can similarly define $[\mathcal{T}', \mathcal{T}]$ for any topos $\mathcal{T}'$.)

The following is an immediate consequence of the comparison theorem for étendues (4.12):

4.14 Corollary. *For any etale topological groupoid G and any CW-complex X there is a canonical bijection*

$$[X, BG] \cong [\mathrm{Sh}(X), \mathcal{B}G] .$$

4.15 Principal bundles. Let G be a topological groupoid. A G-bundle over a space X is a space E equipped with an open surjection $s : E \to X$ and an action by G on the fibers of s. Such an action is given by maps

$$p : E \to G_0 \qquad \text{and} \qquad \mu : G_1 \times_{G_0} E \to E \qquad \mu(g, e) = g \cdot e \,,$$

where μ is defined for points $g \in G_1$ and $e \in E$ such that $d_0(g) = p(e)$; this map μ satisfies the identity

$$s(g \cdot e) = s(e)$$

expressing that the action is fiberwise, as well as the usual identities for a left action:

$$p(g \cdot e) = d_1(g) \,, \qquad g \cdot (h \cdot e) = (g \circ h) \cdot e \,, \qquad i(x) \cdot e = e \,.$$

The G-bundle is called *principal* if the map

$$G_1 \times_{G_0} E \to E \times_X E \qquad (g, e) \mapsto (g \cdot e, e)$$

is a homeomorphism. For two points $e, e' \in s^{-1}(x)$ in the same fiber of a principal bundle, there is a unique arrow $g : p(e) \to p(e')$ such that $g \cdot e = e'$.

These principal G-bundles over X form a category, with as arrows the continuous maps over X which respect the action. Each arrow in this category is an isomorphism. As in [Haefliger (1984)] the set of isomorphism classes of principal G-bundles over X is denoted

$$H^1(X, G) \,.$$

If $f : Y \to X$ is a continuous map, there is an induced functor from principal G-bundles over X to such over Y, sending E to the pullback $f^*(E) = E \times_X Y$ (and structure maps $E \times_X Y \to Y$, $E \times_X Y \to G_0$ and $G_1 \times_{G_0} (E \times_X Y) \to E \times_X Y$ induced from the structure maps of E). This gives an operation

$$f^* : H^1(X, G) \longrightarrow H^1(Y, G)$$

on isomorphism classes.

4.16 Cocycles. For any map between spaces $q : U \to X$ one can construct a topological groupoid $U^{(q)}$ with U as space of objects and $U \times_X U$ as space of arrows; in other words, there is exactly one arrow $y \to y'$ in $U^{(q)}$ iff $y, y' \in U$ are points such that $q(y) = q(y')$. In particular, there is such a groupoid $X^{(\mathrm{id})}$ associated to the identity map on X, and all arrows in this groupoid $X^{(\mathrm{id})}$ are identity arrows. For any map $q : U \to X$ there is a homomorphism of topological groupoids

$$U^{(q)} \to X^{(\mathrm{id})} \,,$$

and this is a weak equivalence iff q is an open surjection. Furthermore, any weak equivalence into $X^{(\mathrm{id})}$ is of this form.

A cocycle on X with values in a topological groupoid G is given by an open surjection $q : U \to X$ and a homomorphism $U^{(q)} \to G$. This is precisely a generalized homomorphism

$$X^{(\mathrm{id})} \xleftarrow{\simeq} U^{(q)} \longrightarrow G \,,$$

as discussed in 2.13. Two cocycles are said to be equivalent if they represent the same generalized homomorphism from $X^{(\mathrm{id})}$ to G. (By the description of the equivalence relation in 2.13, this means that the cocycles have a common refinement, up to conjugation.)

Now suppose that G is (weakly equivalent to) an *etale* groupoid. Noticing that $\mathcal{B}(X^{(\mathrm{id})}) = \mathrm{Sh}(X)$, we then find that by the localization theorem 3.14, there is a bijective correspondence between equivalence classes of cocycles on X and mappings of topoi

$$\mathrm{Sh}(X) \longrightarrow \mathcal{B}G\,.$$

We now sketch that cocycles correspond to principal bundles. First, if E is a principal bundle over X (with structure maps s, p and μ as before) then E gives rise to a cocycle, with $s : E \to X$ as open surjection and with homomorphism $E^{(s)} \to G$ given on objects by the map $p : E \to G_0$ and on arrows by the map $E \times_X E \xrightarrow{\sim} G_1 \times_X E \xrightarrow{\pi_1} G_1$. Conversely, from a cocycle $(q : U \to X, \varphi : U^{(q)} \to G)$ one can construct a principal bundle $E = (G_1 \times_{G_0} U)/\sim$; this is the space of pairs (g, u) where $g : y \to z$ is an arrow in G and $u \in U$ is a point with $\varphi(u) = y$, factored out by the equivalence relation $\sim$ which identifies $(g \circ \varphi(u, u'), u)$ and (g, u'). The action by G on this space E is defined using the composition of G.

In this way, one obtains a bijective correspondence between equivalence classes of cocycles and isomorphism classes of principal bundles. Thus the following theorem is a consequence of the localization theorem 3.14.

4.17 Theorem. *Let X be a topological space, and let G be (weakly equivalent to) an etale topological groupoid. There is natural bijective correspondence*

$$\mathrm{Hom}(\mathrm{Sh}(X), \mathcal{B}G) \cong H^1(X, G)\,.$$

4.18 Concordance. Two principal G-bundles E_0 and E_1 over a space X are said to be *concordant* if there exists a principal bundle D over $X \times [0,1]$ such that pulling back along the inclusions $i_0, i_1 : X \hookrightarrow X \times [0,1]$ gives isomorphisms $i_0^*(D) \cong E_0$ and $i_1^*(D) \cong E_1$. This notion of being concordant defines an equivalence relation on all principal G-bundles over X, and the set of equivalence classes is denoted

$$K_G(X)\,.$$

The following result is a consequence of 4.17 and 4.14.

4.19 Corollary. *Let X be a CW-complex, and let G be a topological groupoid which is (weakly equivalent to) an etale one, as in* 4.17. *There is a natural bijective correspondance*

$$[X, BG] \cong K_G(X)\,.$$

In this sense, the classifying space BG "classifies" concordance classes of principal bundles.

5 Smooth étendues and leaf spaces of foliations.

5.1 The leaf space of a foliation. One of the methods to study and classify foliations is to consider invariants of the "transverse structure", or of the "space of leaves". For a manifold M equipped with a foliation $\mathcal{F}$, the most naive approach would be to define these invariants as invariants of the quotient space $M/\sim$, obtained by identifying any two points which lie on the same leaf. However, it is already quite clear from the elementary examples in §1 that this quotient space $M/\sim$ contains very little information. For the Kronecker foliation (1.3.2) it is an indiscrete space; for the infinite Möbius band (1.3.3) it is the half-line $[0, \infty)$, which is contractible and doesn't reflect the intuition that the foliated Möbius band has a non-trivial foliated double cover which should somehow be classified by the fundamental group of the space of leaves. Thus one is naturally led to consider more general notions of "space" and define the space of leaves as some kind of quotient of M in

such a category of generalized spaces. There are many proposals in this direction that have been studied in the literature. To mention just a few, Grothendieck (SGA IV) proposes the topos $\mathrm{Sh}(M, \mathrm{r})$ of 3.9 as space of leaves; Barre (1973) introduces the notion of Q-variety to study certain types of foliation; Pradines (1979) proposes the more general QF-varieties, and Tapia (1987) studies QF-varieties and their relation to topos theory; Van Est (1984) introduces S-atlases as generalized manifolds in order to define and study the fundamental group of the "space of leaves"; Haefliger (1984) concentrates on invariants of the classifying space $B\,\mathrm{Hol}(M, \mathcal{F})$. Other related sources include Haefliger (1980), Molino (1975), Connes (1982). We will define a quotient "space" of leaves in the following way:

5.2 Definition of $M/\mathcal{F}$. Let $(M, \mathcal{F})$ be a foliated manifold. The *space of leaves* $M/\mathcal{F}$ is defined as

$$M/\mathcal{F} \;=\; \mathcal{B}\,\mathrm{Hol}(M, \mathcal{F})\,.$$

In other words, $M/\mathcal{F}$ is the classifying topos of the holonomy groupoid of M, described in 2.11. As observed above (3.8(ii)), this topos $M/\mathcal{F}$ is in fact an étendue. If T is a complete transversal section for $(M, \mathcal{F})$ then there is an etale topological groupoid $\mathrm{Hol}_T(M, \mathcal{F})$ (see 3.8(ii)) which is weakly equivalent to $\mathrm{Hol}(M, \mathcal{F})$. Therefore $\mathrm{Hol}(M, \mathcal{F})$ and $\mathrm{Hol}_T(M, \mathcal{F})$ define the *same* classifying topos. (More precisely, there is an equivalence of categories $\mathcal{B}\,\mathrm{Hol}_T(M, \mathcal{F}) \cong \mathcal{B}\,\mathrm{Hol}(M, \mathcal{F})$.) In particular, we could equivalently define

$$M/\mathcal{F} \;=\; \mathcal{B}\mathrm{Hol}_T(M, \mathcal{F})$$

for *any* complete transversal section whatsoever.

This definition is compatible with many of the approaches listed above. For example, the comparison theorem 4.12 states that $M/\mathcal{F}$ has the same weak homotopy type as the classifying space $B\,\mathrm{Hol}(M, \mathcal{F})$ (or $B\,\mathrm{Hol}_T(M, \mathcal{F})$) studied by Haefliger. (But $M/\mathcal{F}$ has a natural differentiable structure, as we shall see, whereas $B\,\mathrm{Hol}(M, \mathcal{F})$ does not.) And as remarked in Moerdijk (1991) §4 and in 4.7(iii) above, the fundamental group of the topos $M/\mathcal{F}$ is isomorphic to the fundamental group defined by Van Est (1984) for the S-atlas associated to the foliation $(M, \mathcal{F})$.

5.3 Smooth étendues. There are various equivalent definitions of smooth étendues (or étendues of class C^∞). The one which is close to the original definition 3.5 states that a ringed topos $\mathcal{T}$ is a smooth étendue if there exists an object $E \in \mathcal{T}$ such that $E \to 1$ is epi in $\mathcal{T}$ and $\mathcal{T}/E$ is equivalent (as a ringed topos) to the category $\mathrm{Sh}(M)$ of sheaves on some paracompact Hausdorff C^∞-manifold, equipped with its usual structure sheaf of smooth functions. In line with the Grothendieck-Verdier theorem 3.7 one can also define a smooth étendue as a topos of the form $\mathcal{B}G$ where G is an etale differentiable groupoid (we assume that the space of objects G_0 is paracompact and Hausdorff, but we do not make the same requirement for G_1).

5.4 Smooth maps between étendues and leaf spaces. Let $\mathcal{T} = \mathcal{B}G$ and $\mathcal{T}' = \mathcal{B}H$ be two étendues, associated to topological groupoids G and H which are etale (or weakly equivalent to etale ones). Recall that mappings of topoi

$$\mathcal{B}H \longrightarrow \mathcal{B}G$$

correspond bijectively to equivalence classes of generalized homomorphisms

$$H \xleftarrow{\sim} K \longrightarrow G$$

(see 3.14), and also to isomorphism classes of principal bundles

$$P \in H^1(\mathcal{B}H, G)$$

(see 4.17). If $\mathcal{T}$ and $\mathcal{T}'$ are *smooth* étendues, there is a natural notion of smooth map $f : \mathcal{T} \to \mathcal{T}'$, namely as a map of ringed topoi, or equivalently, as a map which is "covered" by a C^∞-map of manifolds $M \to M'$:

$$\begin{array}{ccccccc} \mathcal{T}' & \longleftarrow & \mathcal{T}/E & \cong & \mathrm{Sh}(M) & & M \\ \downarrow & & \downarrow & & \downarrow & \Longleftrightarrow & \downarrow \\ \mathcal{T}' & \longleftarrow & \mathcal{T}'/E' & \cong & \mathrm{Sh}(M') & & M' . \end{array} \tag{5.4.1}$$

Note that if V is any smooth manifold then $\mathrm{Sh}(V)$ is a smooth étendue. Furthermore, with this definition of smooth map between smooth étendues, there is a bijective correspondence between C^∞-maps of manifolds $V \to W$ and smooth map of étendues $\mathrm{Sh}(V) \to \mathrm{Sh}(W)$. In other words, the category of C^∞-manifolds is a full subcategory of that of C^∞-étendues.

By our earlier results, smooth maps between étendues can equivalently be described as follows. If G and H are differentiable (etale) groupoids representing $\mathcal{T}$ and $\mathcal{T}'$ as above, a smooth map $\mathcal{T} \to \mathcal{T}'$ is represented by a generalized homomorphism

$$G \xleftarrow{\sim} K \longrightarrow H$$

where K is also differentiable and both homomorphisms $K \to G$ and $K \to H$ are differentiable. Or equivalently, a smooth map is represented by a *smooth* principal bundle $P \in H^1(\mathcal{B}H, G)$ (i.e., P is a manifold, all structure maps are C^∞, the map $P \to H_0$ is a submersion, etc.).

In particular, we *derive* an explicit description of smooth maps between leaf spaces of foliated manifolds

$$M/\mathcal{F} \longrightarrow M'/\mathcal{F}' .$$

This is *exactly* the notion of map between leaf spaces used by Connes (1982), Hilsum-Skandalis (1987). In other words, their maps are precisely smooth mappings between classifying topoi.

5.5 The tangent bundle of a smooth étendue. Let $\mathcal{T}$ be a smooth étendue, and write $\mathcal{T} = \mathcal{B}G$ for a suitable etale differentiable groupoid G. By applying the tangent bundle functor T to G, we obtain another etale differentiable groupoid $T(G)$, with $T(G_0)$ as space of objects and $T(G_1)$ as space of arrows. Furthermore, there is an evident projection homomorphism of differentiable groupoids

$$T(G) \xrightarrow{\pi} G .$$

Thus one obtains an étendue $\mathcal{B}(TG)$ and a map of étendues $\pi : \mathcal{B}(TG) \to \mathcal{B}G = \mathcal{T}$. By definition the tangent bundle $T(\mathcal{T})$ of the étendue $\mathcal{T}$ is $\mathcal{B}(TG)$. It can be shown that this construction does not depend on the groupoid G chosen to represent $\mathcal{T}$, so the notation $T(\mathcal{T})$ is justified.

The construction extends the usual construction of the tangent bundle of a manifold, in the sense that if the étendue $\mathcal{T} = \mathrm{Sh}(M)$ is the topos of sheaves on a manifold M then $T(\mathrm{Sh}(M)) = \mathrm{Sh}(TM)$.

Using the tangent bundle, one can introduce the usual notions of differentiable geometry such as submersions (cf. 5.7 below), immersions, vectorfields, differential forms, Riemannian metrics, etc. in a straight forward way. This is closely related to the transversal structures defined for foliated manifolds. For example, with the evident notion of Riemannian étendue, we find that a foliation $(M,\mathcal{F})$ is Riemannian (see e.g. Reinhart (1983), Molino (1988)) iff $M/\mathcal{F}$ (as defined in 5.2) is a Riemannian étendue. A similar correspondence holds for differential forms, cf. 5.6.1 below.

5.6 The De Rham complex of a smooth étendue. Let $\mathcal{T} = \mathcal{B}G$ be an étendue, represented by an etale differentiable groupoid G, as before. Let Ω^n be the sheaf of germs of differentiable n-forms on G_0. There is an action of G on Ω^n, defined using the germ of a diffeomorphism $\tilde{g} : V_x \to W_y$ for each arrow $g : x \to y$ of G, as in 2.8. Thus for a germ ω_y at y of an n-form ω defined on a neighborhood of y, the arrow g acts as

$$\omega_y \cdot g \;=\; \tilde{g}^*(\omega)_x \,.$$

This makes Ω^n into a G-sheaf, i.e. an object of $\mathcal{T} = \mathcal{B}G$. In this way, one obtains a resolution in $\mathcal{B}G$

$$0 \to \Delta(\mathbf{R}) \to \Omega^0 \to \Omega^1 \to \ldots\ldots \tag{5.6.1}$$

of the constant G-sheaf $\Delta(\mathbf{R})$. This resolution does not depend on the groupoid G chosen to represent $\mathcal{T}$. Applying the global sections functor $\Gamma : \mathcal{T} \to \mathbf{Sets}$ to (5.6.1), one obtains a complex of abelian groups (vector spaces)

$$\Gamma\Omega^0 \to \Gamma\Omega^1 \to \ldots \,. \tag{5.6.2}$$

The elements of $\Gamma\Omega^n$ are the *G-invariant* differential n-forms on G_0, i.e. forms ω on G_0 with the property that for any arrow $g : x \to y$ in G,

$$\omega_y \cdot g = \omega_x \,.$$

By definition, the *de Rham cohomology* of $\mathcal{T}$ is the cohomology of the complex (5.6.2):

$$H^n_{dR}(\mathcal{T}) \;:=\; H^n(\Gamma\Omega^*) \,.$$

This definition agrees with the usual one in case $\mathcal{T} = \mathrm{Sh}(M)$ is the étendue of sheaves on a manifold M, in the sense that

$$H^n_{dR}(\mathrm{Sh}(M)) \;\cong\; H^n_{dR}(M) \,.$$

For a foliated manifold $(M,\mathcal{F})$, one thus obtains the de Rham cohomology groups

$$H^n_{dR}(M/\mathcal{F})$$

of the associated space (i.e., étendue) of leaves.

A *basic differential n-form* on $(M,\mathcal{F})$ is a form ω on M which depends only on the transversal structure of the foliation. This is expressed by saying that (i) ω vanishes for any n-tuple of tangent vectors with the property that (at least) one of them lies in a leaf, and (ii) ω "does not change" when one compares ω_x and ω_y for two points x and y which are infinitesimally close and are on the same leaf. [This can be expressed formally in terms of the interior product i and the Lie derivative θ, by the equations

$$i_X(\omega) = 0 \quad \text{and} \quad \theta_X(\omega) = 0$$

for any (local) vector field X in a leaf L; see e.g. Tondeur(1988), p. 120.] These basic forms constitute a subcomplex Ω^*_{basic} of the de Rham complex of M. The *basic de Rham cohomology* groups $H^n_{\text{basic}}(M, \mathcal{F})$ of the foliation are defined as the cohomology groups of this complex.

5.6.3 Proposition. *For any foliated manifold, there is a canonical isomorphism*

$$H^*_{\text{basic}}(M, \mathcal{F}) \cong H^*_{dR}(M/\mathcal{F})$$

between the basic de Rham cohomology of M and the ordinary de Rham cohomology of the leaf space $M/\mathcal{F}$.

The *de Rham theorem* for differentiable manifolds states that $H^*_{dR}(M)$ is isomorphic to $H^*(M, \Delta\mathbf{R})$, the sheaf cohomology with coefficients in the constant sheaf $\mathbf{R}$. This theorem does not extend to foliations. For example, consider the Kronecker foliation $(M, \mathcal{F})$ of the torus, with slope α so that all leaves are dense. Then $M/\mathcal{F}$ is (equivalent to) the étendue

$$\mathrm{Sh}(S^1, \mathbf{Z})$$

defined by the action of $\mathbf{Z}$ on the circle S^1 given by rotation along an angle α. (This is an étendue of the form 3.1(iv), 3.8(iii)). It is not difficult to show that

$$\pi_1 \mathrm{Sh}(S^1, \mathbf{Z}) \cong \mathbf{Z} \oplus \mathbf{Z}$$

so that, by the Hurewicz formula (4.8)

$$H^1(\mathrm{Sh}(S^1, \mathbf{Z}), \Delta\mathbf{R}) \cong \mathbf{R} \oplus \mathbf{R}\,.$$

On the other hand

$$H^1_{dR}(\mathrm{Sh}(S^1, \mathbf{Z}), \Delta\mathbf{R}) \cong \mathbf{R}\,.$$

[Indeed, any invariant 1-form is of the form $f(t)dt$ where $f : S^1 \to \mathbf{R}$ is constant, and no such form (except $f = 0$) is a boundary dg for a $\mathbf{Z}$-invariant function $S^1 \to \mathbf{R}$.]

5.7 Submersions. A map $f : \mathcal{T} \to \mathcal{T}'$ between C^∞-étendues is a submersion if it is covered (as in diagram 5.4.1) by a submersion of manifolds $M \to M'$.

5.8 Classifying foliations. Recall from 2.6 and 3.8(i) the Haefliger groupoid Γ^q and its associated étendue $\mathcal{B}\Gamma^q$.

5.8.1 Theorem. (Kock-Moerdijk (unpublished)) *Let M be an C^∞-manifold. There is a bijective correspondence between foliations on M of codimension q and (isomorphism classes of) submersions between étendues* $\mathrm{Sh}(M) \to \mathcal{B}\Gamma^q$.

5.8.2 Remarks. (a) The proof of 5.8.1 is not difficult, and can be outlined as follows. In one direction, suppose a foliation on M is given in the form 1.4.1. Using the notation of 4.16, the cover $M = \bigcup U_i$ gives an etale surjection $s : U = \sum U_i \to M$ and hence a weak equivalence $U^{(s)} \to M^{(\mathrm{id})}$ of differentiable groupoids. Furthermore, the $q_i : U_i \to \mathbf{R}^q$ and the h_{ij} of 1.4.1 give a homomorphism $U^{(s)} \to \Gamma^q$. Then one obtains a generalized homomorphism from $M^{(\mathrm{id})}$ to Γ^q, hence a map $\mathrm{Sh}(M) \to \mathcal{B}\Gamma^q$. This map is a submersion.

In the other direction, a given submersion $\mathrm{Sh}(M) \xrightarrow{s} \mathcal{B}\Gamma^q$ induces a map of tangent bundles $T(M) \twoheadrightarrow s^*T(\mathcal{B}\Gamma^q)$, the kernel of which is an integrable subbundle of $T(M)$ of codimension q. This gives the required foliation on M by 1.4.2.

(b) Although the classifying space $B\Gamma^q$ is weakly equivalent to the classifying topos $\mathcal{B}\Gamma^q$ (see 4.12), the space $B\Gamma^q$ does not have any C^∞-structure, and it is hard to capture "submersions" $M \to B\Gamma^q$ and give a purely topological classification theorem analogous to 5.8.1 –see Haefliger (1972).

5.9 Construction of the holonomy groupoid. (outline) We will end these lectures with a brief indication of how topos-theory can be used to give an elegant construction of the holonomy groupoid $\mathrm{Hol}(M, \mathcal{F})$ (see 2.11) of a foliation $(M, \mathcal{F})$. From this construction it will be immediately obvious that $\mathrm{Hol}(M, \mathcal{F})$ is a differentiable groupoid.

Recall first from 4.16 that any map $q : U \to X$ gives rise to a groupoid $U^{(q)}$ and a homomorphism $U^{(q)} \to X^{(\mathrm{id})}$, inducing a map of étendues $\mathcal{B}U^{(q)} \to \mathrm{Sh}(X)$. Note that $U^{(q)}$ is a differentiable groupoid if U and X are manifolds and q is a submersion. Exactly the same construction applies when q is a map into a smooth étendue. More precisely, if U is a manifold and

$$q \ : \ \mathrm{Sh}(U) \longrightarrow \mathcal{T}$$

is a submersion, then $\mathrm{Sh}(U) \times_{\mathcal{T}} \mathrm{Sh}(U) = \mathrm{Sh}(H)$ for some manifold H, and the projections $\mathrm{Sh}(U) \times_{\mathcal{T}} \mathrm{Sh}(U) \rightrightarrows \mathrm{Sh}(U)$ correspond to C^∞-maps $d_0, d_1 : H \rightrightarrows U$, which are the domain and codomain maps of a differentiable groupoid $U^{(q)}$.

The second ingredient of a general nature, used in the construction of $\mathrm{Hol}(M, \mathcal{F})$, is the following. If $f : M \to N$ is a submersion between manifolds, then one can construct the "fiberwise set of connected components", to factor f as

$$\begin{array}{ccc} M & \xrightarrow{p} & \pi_0(f) \\ & \searrow^{f} & \downarrow e \\ & & N \end{array}$$

where e is a local diffeomorphism (a sheaf on N) and p is a submersion with connected fibers. Exactly the same construction applies to a submersion $f : \mathcal{T} \to \mathcal{T}'$ of smooth étendues: There is a uniquely determined factorization $f = e \circ p$

$$\begin{array}{ccc} \mathcal{T} & \xrightarrow{p} & \pi_0(f) \\ & \searrow^{f} & \downarrow e \\ & & \mathcal{T}' \end{array}$$

where p is a submersion with connected fibers (i.e., f^* is full and faithful) and e is a local diffeomorphism. In fact $\pi_0(f)$ is a smooth étendue of the form $\mathcal{T}'/E$ for a suitable object E of $\mathcal{T}$ and $\pi_0(f) \to \mathcal{T}'$ is the canonical map $\mathcal{T}'/E \to \mathcal{T}'$.

Now consider a foliated manifold $(M, \mathcal{F})$ of codimension q, and its "classifying submersion" (Theorem 5.8.1)

$$c_{\mathcal{F}} : \mathrm{Sh}(M) \longrightarrow \mathcal{B}\Gamma^q \ .$$

Factor $c_{\mathcal{F}}$ as a submersion with connected fibers followed by a local diffeomorphism

$$\mathrm{Sh}(M) \xrightarrow{p_{\mathcal{F}}} \pi_0(c_{\mathcal{F}}) \xrightarrow{e_{\mathcal{F}}} \mathcal{B}\Gamma^q \ ,$$

Now construct the differentiable groupoid $M^{(p_{\mathcal{F}})}$, with M as space of objects, and space of arrows H uniquely defined by

$$\mathrm{Sh}(H) \;=\; \mathrm{Sh}(M) \times_{\pi_0(c_{\mathcal{F}})} \mathrm{Sh}(M)\ .$$

5.9.1 **Theorem.** *The differentiable groupoid $H \rightrightarrows M$ thus constructed is precisely the holonomy groupoid of $(M, \mathcal{F})$.*

References

[SGA4] M. Artin, A. Grothendieck, J.-L. Verdier, *Théorie de topos et cohomologie étale des schémas*, Springer LNM 269 and 270 (1972).

M. Artin, B. Mazur, *Etale Homotopy*, Springer LNM 100 (1969).

R. Barre, De quelques aspects de la théorie des Q-variétés différentielles et analytiques, Ann. Inst. Fourier, Grenoble, 23 (1973), pp. 227-312.

M. Barr, R. Diaconescu, On locally simply connected toposes and their fundamental groups, Cahiers Top. Géom. Diff. 22 (1981), pp. 301-314.

R. Bott, Characteristic classes and foliations, in: Springer LNM 279 (1972), pp. 1-94.

K. Brown, *Cohomology of groups*, Springer-Verlag, Berlin, etc., 1982.

M. Bunge, An application of descent to a classification theorem for toposes, Math. Proc. Cambridge Phil. Soc. 107 (1990), pp. 59-79.

C. Camacho, A. Neto, *Geometric Theory of Foliations*, Birkhäuser, Boston, 1985.

A. Connes, A survey of foliations and operator algebras, Proc. Symp. Pure Math. **38**, Part I (1982), pp. 521-628.

J. L. Dupont, *Curvature and Characteristic Classes*, Springer LNM 640 (1978).

W. T. van Est, Rapport sur les S-atlas, Astérisque **116** (1984), pp. 235-292.

F. Frobenius, Ueber das Pfaffsche Problem, J. reine angew. Math. **82** (1877), p. 267-282.

D. Fuks, Foliations, J. Soviet Math. 18 (1982), pp. 255-291.

P. Gabriel, M. Zisman, *Calculus of Fractions and Homotopy Theory*, Springer-Verlag, Berlin etc., 1967.

R. Godement, *Topologie algébrique et théorie des faisceaux*, Hermann, Paris, 1958.

C. Godbillon, Feuilletages, Birkhäuser, Basel, 1991.

[SGA1] A. Grothendieck, Revêtements étales et groupe fondamental, Springer LNM 224 (1971).

A. Haefliger, Structures feuilletées et cohomologie à valeur dans un faisceau de groupoïdes, Comm. Math. Helv. **32** (1958), pp. 248-329.

A. Haefliger, Homotopy and integrability, in Springer LNM 197 (1971), pp. 133-163.

A. Haefliger, Some remarks on foliations with minimal leaves, J. Diff. Geom. **15** (1980), pp. 269-284.

A. Haefliger, Groupoïdes d'holonomie et classifiants, Astérix **116** (1984), pp. 70-97.

G. Hector, U. Hirsch, Introduction to the Geometry of Foliations (two volumes), Vieweg, Braunschweig, 1981 and 1983.

M. Hilsum, G. Skandalis, Morphismes K-orientés d'espace de feuilles et fonctorialité en théorie de Kasparov, Ann. Sci. Ec. Norm. Sup. 20 (1987), pp. 325-390.

A. Joyal, I. Moerdijk, Toposes as homotopy groupoids, Adv. in Math. 80 (1990), pp. 22-38.

A. Joyal, M. Tierney, An extension of the Galois theory of Grothendieck, Memoirs A.M.S. 309 (1984).

J. Kennison, Pro-group actions and fundamental progroups, J. Pure and Appl. Alg. 66 (1990), pp. 185-218.

A. Kock, I. Moerdijk, Local equivalence relations and their sheaves, Aarhus preprint no.19 (1991).

A. Kock, I. Moerdijk, Every étendue comes from a local equivalence relation, J. Pure and Appl. Alg. 82 (1992), pp. 155-174.

A. Kock, I. Moerdijk, Topological spaces with local equivalence relations; in preparation.

H. B. Lawson, Lectures on the quantitative theory of foliations, CBMS Regional Conf. Series 27 (1977).

R. Lickorish, A presentation of orientable combinatorial 3-manifolds, Ann. Math. 76 (1962), pp. 531-540.

R. Lickorish, A foliation for 3-manifolds, Ann. Math. 82 (1965), pp. 414-420.

S. Mac Lane, *Categories for the Working Mathematician*, Springer-Verlag, Berlin etc., 1971.

S. Mac Lane, I. Moerdijk, *Sheaves in Geometry and Logic*, Springer-Verlag, Berlin etc., 1992.

I. Moerdijk, The classifying topos of a continuous groupoid, I, Transactions A.M.S. 310 (1988a), pp. 629-668.

I. Moerdijk, Toposes and groupoids, Springer LNM 1348 (1988b), pp. 280-298.

I. Moerdijk, Prodiscrete groups and Galois toposes, Indag. Math. 51 (1989), pp. 219-234.

I. Moerdijk, Classifying toposes and foliations, Ann. Inst. Fourier 41, Grenoble (1991), pp. 189-209.

P. Molino, Sur la géometrie transverse des feuilletages, Ann. Inst. Fourier 25 (1975), pp. 279-284.

P. Molino, *Riemannian Foliations*, Birkhäuser, Boston, 1988.

J. Philips, The holonomic imperative and the homotopy groupoid of a foliated manifold, Rocky Mountains Journ. Math. 17 (1987), pp. 151-165.

J. Pradines, Wouafo-Kamga, J., La catégorie des QF-variétés, C.R. Acad. Sci. Paris 288 (1979), pp. 717-719.

Pronk, D., Stacks, étendues and bicategories of fractions, in preparation.

B. Reinhart, *Differential Geometry of Foliations*, Springer-Verlag, Berlin etc., 1983.

G. Segal, Classifying spaces and spectral sequences. Inst. Hautes Et. Sci., Publ. Math. 34 (1968), pp. 105-112.

J. P. Serre, Lie groups and Lie algebras, Harvard University, 1964. (Reprinted as Springer LNM 1500 (1992).)

Ph. Tondeur, *Foliations on Riemannian Manifolds*, Springer-Verlag, Berlin etc., 1988.

J. Tapia, Sur la cohomologie de certains espaces d'orbites, thèse, Univ. P. Sabatier, Toulouse (1987).

H. Winkelnkemper, The graph of a foliation, Ann. Global Anal. Geom. 1 (1983),pp.

Rev. Academia de Ciencias. Zaragoza. 48 (1993)

CHARACTERIZATION OF COMPLEX FILIFORM LIE ALGEBRAS OF DIMENSION 8 ACCORDING TO WHETHER THEY ARE OR NOT DERIVED FROM OTHERS

By

F.J. ECHARTE REULA[1], J.R. GOMEZ MARTIN[2] y J. NUÑEZ VALDES[1].

[1] Universidad de Sevilla. Facultad de Matemáticas. Dpto de Algebra, Computación, Geometría y Topología. C/ Tarfia s/n. 41012 Sevilla. España.

[2] Universidad de Sevilla. Facultad de Informática y Estadística. Dpto de Matemática Aplicada. C/ Tarfia s/n. 41012 Sevilla, España.

AMS (1980) Subject Classification: 17 B 30

Abstract.- In this paper, we characterize those Complex Filiform Lie Algebras of dimension 8 which are derived from other Solvable Lie Algebras of higher dimension. This result and the previous one given in ([8]) allow us to find a complete list of Characteristically Nilpotent Filiform Lie Algebras of dimension 8.

1.- Introduction and Notations.

There does not exist, at present, any classification of complex Nilpotent Lie Algebras (NLA) of dimension greater than 7. Goze and Ancochea, by the introduction of a new invariant which they call the "characteristic sequence", which corresponds to the maximal dimensions of Jordan blocks of a certain nilpotent matrix, obtained the classification of complex Filiform Lie Algebras (FLA) of dimension 8 ([1]). These FLA, as it is known, are a subset of NLA.

The authors are supported by the project PAICYT of the Junta de Andalucía. España (1990).

However, the importance of these authors' work is not only to have obtained this classification, but, above all, to have devised techniques which can be applied in any dimension, which, apart from the fact of having permitted Gómez Martín to classify FLA of dimension 9 ([5]), totally solved the problem of this classification of FLA of greater dimension, although, of course, this requires hard and complicated calculations which are impossible without the use of a computer.

Therefore, although the classification of Lie Algebras (LA) of dimension greater than 7 has not been obtained yet, except for FLA, the main problem which is now considered is the search for new results, rather than the classification itself, such as the attainment of new invariants or the description of the irreducible components of varieties of LA. For this purpose, Valeiras proved recently that the variety of NLA of dimension 8 has 8 irreducible components and that only the two first of them meet the open set of FLA ([9]).

The main purpose of this paper is, therefore, starting from the classification of FLA of dimension 8 by Goze and Ancochea, to characterize two groups of these LA, according to whether or not they are derived from other Solvable Lie Algebras (SLA) of higher dimension. This characterization we propose could be interesting in the sense of knowing which of these FLA are also Characteristically Nilpotent Lie Algebras (CNLA) since, as we previously proved in an earlier paper ([8]), if a FLA is not derived from any LA then it is a CNLA. This theorem will allow us to write the list of Characteristically Nilpotent Filiform Lie Algebras (CNFLA) of dimension 8.

A complex FLA may be derived from an SLA of higher dimension. In ([8]) we proved the following:

(1.1) Theorem: A complex FLA of dimension n is either derived from a SLA of dimension $n+1$ or not derived from any LA.

From now on, we write $(A,B,C) = 0$ to represent the Jacobi identity

$$[[A,B],C] + [[B,C],A] + [[C,A],B] = 0$$

and we denote by $\mathfrak{N}$ a FLA of dimension n ([8]), that is, a complex NLA admitting a basis

(1) $$\{X_1, X_2, \ldots, X_n\}$$

such that $X_1 \notin [\mathfrak{N}, \mathfrak{N}]$,

(2) $$[X_1, X_2] = 0$$

and

(3) $$[X_1, X_j] = X_{j-1} \qquad 3 \le j \le n$$

Moreover, since $[X_s, X_j] = 0 \quad 1 < j < n$ and $[X_s, X_n] = a\ X_2$ ma $\in C$), if we consider the change of basis given by

$$X_j' = X_j \quad (1 \le j \le n-1) \quad ; \quad X_n' = X_n + a\ X_1$$

then it results that $[X_s', X_j'] = 0$ $(1 < j \le n)$. So, a basis (1) can be always chosen in a suitable way verifying

(4) $$[X_s, X_j] = 0 \qquad 1 < j \le n$$

Moreover, $\mathfrak{L}$ will denote a complex SLA of dimension n+1 such that

(5) $$\{X_1, \ldots, X_n, U\}$$

is a basis, where $\{X_1, X_2, \ldots, X_n\}$ is the basis (1) mentioned above, and U a derivation of $\mathfrak{M}$ such that

(6) $$[X_j, U] = \sum_{h=1}^{n} a_{jh}\ X_h$$

with $(X_j, X_h, U) = 0, \quad 1 \le j,h \le n$.

The FLA $\mathfrak{M}$ is said to be derived from the SLA $\mathfrak{L}$ if $[\mathfrak{L},\mathfrak{L}] \equiv \mathfrak{M}$, where $[\mathfrak{L},\mathfrak{L}]$ represents any linear combination of all brackets among fields of the basis (1).

Characteristically Nilpotent Lie Algebras are those LA in which all their derivations are nilpotent. The first examples of CNLA given in the literature were of LA that are not derived from any LA, until Luks gave an example of a CNLA which is derived from a LA of dimension 18 ([7]). However, it is very easy to give examples of Nilpotent Lie Algebras (NLA) which are derived and not CNLA.

As we mentioned above, in ([8]) the following two theorems are proved:

(1.2) Theorem A necessary and sufficient condition for a FLA $\mathfrak{M}$ to be derived from the SLA $\mathfrak{L}$ is that $a_{11} \neq 0$.

(1.3) Theorem A complex FLA $\mathfrak{M}$ is a CNLA if and only if $\mathfrak{M}$ is not derived from any LA.

So, to separate FLA of dimension 8 into two groups according to whether or not they are derived from another SLA (Theorem (1.1)) we will use a method previously indicated by us for the case of NLA of dimension 7 ([4]). This proceedure, which we called the "method of determination of the vanishing of the coefficient a_{11}", can be applied to FLA. It consists of a set of iterative calculations based on all

possible Jacobi identities of the kind $(X_i, X_j, U) = 0$ with $1 \leq i,j \leq 8$, from which we can determine whether the coefficient a_{11} vanishes or not. This in turn determines, by theorem (1.2), whether the FLA $\mathfrak{M}$ is derived from the SLA $\mathfrak{L}$ with $\{X_1,..,X_8,U\}$ as a basis.

These iterative steps are the following:

1.- Using the Jacobi identity $(X_1, X_{h_1}, U) = 0$ with $h_1 > 1$, deriving from such identity the set of equations obtained by setting to zero the coefficient of each field X_j.

2.- Using each equation obtained in this way to express in terms of the others the coefficient a_{ij} whose pair of subindices is the greater (in lexicographic order). (For instance, from the equation $a_{22} - a_{11} - a_{44} - a_{17} = 0$ we will have $a_{44} = a_{22} - a_{11} - a_{17}$).

3.- Replacing in (6) the coefficients a_{ij} with their corresponding values obtained previously in step 2.

4.- Repeating the former three steps with U, X_2 and X_{h_2} with $h_2 > 2$ in the first place. Secondly, with U, X_3 and X_{h_3} with $h_3 > 3$ and so on, until terminating this proceeding with U, X_{n-1} and X_n.

5.- Finally, observing if in the new expressions (6) obtained in this way the coefficient a_{11} appears explicity or not; this will indicate, according to theorem (1.2) whether the FLA $\mathfrak{M}$ which we are studying is derived from another SLA $\mathfrak{L}$ of dimension one more than the dimension of $\mathfrak{M}$.

In the case that the FLA $\mathfrak{M}$ is derived from another SLA of dimension one greater than the dimension of $\mathfrak{M}$, we can obtain the simplest SLA $\mathfrak{L}$ by taking $a_{11} = 1$ and the rest of the coefficients $a_{ij} = 0$ $(i,j \neq 1)$. Moreover, we can check that in FLA of dimension ≤ 9, either directly or by suitable changes of base, it is always possible to get

(7) $$[X_h, U] = a_h X_h \quad \text{with } h = 1, 2, \ldots, n$$

(although this is not proved in the general case of dimension $\mathfrak{M} = n$).

2.- Division of complex FLA of dimension 8 into two groups depending on whether they are or not derived from other LA.

We separate FLA of dimension 8 into two groups depending on whether they are derived from another SLA, starting from the classification of these FLA by Goze and Ancochea ([1]). This classification (which is indicated in an Appendix at the end of this paper) contains a list of 13 isolated FLA and 7 one-parameter families of FLA.

By the method mentioned above, we find that 6 of the 13 FLA and 2 of the 7 families are derived from SLA of dimension 8 whereas the other 7 FLA and the 5 families are not.

So we obtain the following

Theorem 1.- The complex FLA of dimension 8 with laws μ_8^1, μ_8^2, μ_8^3, μ_8^4, $\mu_8^{5,\alpha}$, μ_8^6, $\mu_8^{9,\alpha}$, $\mu_8^{10,\alpha}$, μ_8^{11}, $\mu_8^{13,\alpha}$, $\mu_8^{15,\alpha}$ and μ_8^{17} are not derived from any other LA.

Proof:

By appliying the method mentioned above to each of them we obtain the following results:

(2.1) The FLA of dimension 8 with law μ_8^1 is not derived from any LA, since:

$$
\begin{aligned}
[X_1,U] &= a_{12} X_2 + a_{13} X_3 + a_{14} X_4 + a_{15} X_5 + a_{16} X_6 + a_{17} X_7 \\
[X_2,U] &= 0 \\
[X_3,U] &= a_{32} X_2 \\
[X_4,U] &= (a_{16} + a_{17}) X_2 + (a_{32} + a_{17}) X_3 \\
[X_5,U] &= a_{52} X_2 + 3/5\, a_{17} X_3 + (a_{32} + a_{17}) X_4 \\
[X_6,U] &= (3/5\, a_{52} + a_{14} + 3/5\, a_{15} + 6/25\, a_{16}) X_2 + \\
&\quad + (a_{52} + a_{15}) X_3 + 1/5\, a_{17} X_4 + (a_{32} + a_{17}) X_5 \\
[X_7,U] &= a_{72} X_2 + (3/5\, a_{52} + a_{15} + 6/25\, a_{16}) X_3 + \\
&\quad + (a_{52} + a_{15} + 2/5\, a_{16}) X_4 + 1/5\, a_{17} X_5 + (a_{32} + a_{17}) X_6 \\
[X_8,U] &= a_{82} X_2 + (a_{72} - a_{14}) X_3 + (3/5\, a_{52} - a_{14} + 2/5\, a_{15} + 6/25\, a_{16}) X_4 + \\
&\quad + (a_{52} + 1/5\, a_{16}) X_5 - a_{16} X_6 + a_{32} X_7
\end{aligned}
$$

(2.2) The FLA of dimension 8 with law μ_8^2 is not derived from any LA, since:

$$[X_1,U] = a_{12} X_2 + a_{13} X_3 + a_{14} X_4 + a_{15} X_5 + a_{16} X_6 + a_{17} X_7$$
$$[X_2,U] = 0$$
$$[X_3,U] = a_{32} X_2$$
$$[X_4,U] = a_{16} X_2 + (a_{32} + a_{17}) X_3$$
$$[X_5,U] = a_{52} X_2 + (a_{32} + a_{17}) X_4$$
$$[X_6,U] = (a_{17} + a_{16} + a_{14}) X_2 + (a_{52} + a_{17} + a_{15}) X_3 + (a_{32} + a_{17}) X_5$$
$$[X_7,U] = a_{72} X_2 + a_{17} X_3 + (a_{52} + a_{15} + a_{17}) X_4 + (a_{32} + a_{17}) X_6$$
$$[X_8,U] = a_{82} X_2 + (a_{72} - a_{16}) X_3 - (a_{16} - a_{14}) X_4 + a_{52} X_5 - a_{16} X_6 + a_{32} X_7$$

(2.3) The FLA of dimension 8 with law μ_8^3 is not derived from any LA, since:

$$[X_1,U] = a_{12} X_2 + a_{13} X_3 + a_{14} X_4 + a_{15} X_5 + a_{16} X_6 + a_{17} X_7$$
$$[X_2,U] = 0$$
$$[X_3,U] = a_{32} X_2$$
$$[X_4,U] = a_{16} X_2 + (a_{32} + a_{17}) X_3$$
$$[X_5,U] = a_{52} X_2 + (a_{32} + a_{17}) X_4$$
$$[X_6,U] = (a_{17} + a_{14}) X_2 + (a_{52} + a_{15}) X_3 + (a_{32} + a_{17}) X_5$$
$$[X_7,U] = a_{72} X_2 + a_{17} X_3 + (a_{52} + a_{15}) X_4 + (a_{32} + a_{17}) X_6$$
$$[X_8,U] = a_{82} X_2 + (a_{72} - a_{16}) X_3 - a_{14} X_4 + a_{52} X_5 - a_{16} X_6 + a_{32} X_7$$

(2.4) The FLA of dimension 8 with law μ_8^4 is not derived from any LA, since:

$$[X_1,U] = a_{12} X_2 + a_{13} X_3 + a_{14} X_4 + a_{15} X_5 + a_{16} X_6 + a_{17} X_7$$
$$[X_2,U] = 0$$
$$[X_3,U] = a_{32} X_2$$
$$[X_4,U] = a_{16} X_2 + (a_{32} + a_{17}) X_3$$
$$[X_5,U] = a_{52} X_2 + (a_{32} + a_{17}) X_4$$
$$[X_6,U] = (a_{16} + a_{14}) X_2 + (a_{52} + a_{15} + a_{17}) X_3 + (a_{32} + a_{17}) X_5$$
$$[X_7,U] = a_{72} X_2 + (a_{52} + a_{15} + a_{17}) X_4 + (a_{32} + a_{17}) X_6$$
$$[X_8,U] = a_{82} X_2 + a_{72} X_3 - (a_{14} + a_{16}) X_4 + a_{52} X_5 - a_{16} X_6 + a_{32} X_7$$

(2.5) The FLA of dimension 8 with law $\mu_8^{\sigma,\alpha}$ is not derived from any LA, since:

$$[X_1,U] = a_{12} X_2 + a_{13} X_3 + a_{14} X_4 + a_{15} X_5 + a_{16} X_6 + a_{17} X_7$$
$$[X_2,U] = 0$$
$$[X_3,U] = a_{32} X_2$$
$$[X_4,U] = \alpha\, a_{17} X_2 + a_{32} X_3$$
$$[X_5,U] = (a_{16} + a_{17}) X_2 + (1 + \alpha)\, a_{17} X_3 + a_{32} X_4$$
$$[X_6,U] = a_{62} X_2 + (a_{16} + a_{17}) X_3 + (2 + \alpha)\, a_{17} X_4 + a_{32} X_5$$

$$[X_7,U] = a_{72}X_2 + (a_{62} - a_{15})X_3 + a_{17}X_4 + (2+\alpha)a_{17}X_5 + a_{32}X_6$$
$$[X_8,U] = a_{82}X_2 + (a_{72} - \alpha a_{14} - a_{15})X_3 + (a_{62} - (2+\alpha)a_{15} - a_{16})X_4 - (2+\alpha)a_{16}X_5 + a_{32}X_7$$

(2.6) The FLA of dimension 8 with law μ_8^8 is not derived from any LA, since:

$$[X_1,U] = a_{12}X_2 + a_{13}X_3 + a_{14}X_4 + a_{15}X_5 + a_{16}X_6 + a_{17}X_7$$
$$[X_2,U] = 0$$
$$[X_3,U] = a_{32}X_2$$
$$[X_4,U] = -a_{17}X_2 + a_{32}X_3$$
$$[X_5,U] = a_{16}X_2 + a_{32}X_4$$
$$[X_6,U] = a_{62}X_2 + (a_{16} + a_{17})X_3 + a_{17}X_4 + a_{32}X_5$$
$$[X_7,U] = a_{72}X_2 + (a_{62} - a_{15} - a_{16})X_3 + a_{17}X_4 + a_{17}X_5 + a_{32}X_6$$
$$[X_8,U] = a_{82}X_2 + (a_{72} + a_{14})X_3 + (a_{62} - a_{15} - 2a_{16})X_4 - a_{16}X_5 + a_{32}X_7$$

(2.7) The FLA of dimension 8 with law $\mu_8^{9,\alpha}$ is not derived from any LA, since:

$$[X_1,U] = a_{12}X_2 + a_{13}X_3 + a_{14}X_4 + a_{15}X_5 + a_{16}X_6 + a_{17}X_7$$
$$[X_2,U] = 0$$
$$[X_3,U] = a_{32}X_2$$
$$[X_4,U] = a_{17}X_2 + a_{32}X_3$$
$$[X_5,U] = a_{52}X_2 + a_{17}X_3 + a_{32}X_4$$
$$[X_6,U] = a_{62}X_2 + (a_{52} + a_{17})X_3 + a_{17}X_4 + a_{32}X_5$$
$$[X_7,U] = a_{72}X_2 + (a_{62} - a_{16})X_3 + (a_{52} + a_{17})X_4 + a_{17}X_5 + a_{32}X_6$$
$$[X_8,U] = a_{82}X_2 + (a_{72} - a_{14} - \alpha a_{16})X_3 + (a_{62} - a_{15} - 2a_{16} - \alpha a_{17})X_4 + (a_{52} - a_{16})X_5 + a_{32}X_7$$

(2.8) The FLA of dimension 8 with law $\mu_8^{10,\alpha}$ is not derived from any LA, since:

$$[X_1,U] = a_{12}X_2 + a_{13}X_3 + a_{14}X_4 + a_{15}X_5 + a_{16}X_6 + a_{17}X_7$$
$$[X_2,U] = 0$$
$$[X_3,U] = a_{32}X_2$$
$$[X_4,U] = a_{42}X_2 + a_{32}X_3$$
$$[X_5,U] = a_{52}X_2 + a_{42}X_3 + a_{32}X_4$$
$$[X_6,U] = a_{62}X_2 + a_{52}X_3 + a_{42}X_4 + a_{32}X_5$$
$$[X_7,U] = a_{72}X_2 + a_{62}X_3 + a_{52}X_4 + a_{42}X_5 + a_{32}X_6$$
$$[X_8,U] = a_{82}X_2 + (a_{72} - \alpha a_{17} - a_{16} - a_{14})X_3 + (a_{62} - a_{17} - a_{15})X_4 + (a_{52} - a_{16})X_5 + (a_{42} - a_{17})X_6 + a_{32}X_7$$

(2.9) The FLA of dimension 8 with law μ_8^{11} is not derived from any LA, since:

$$[X_1,U] = a_{12}X_2 + a_{13}X_3 + a_{14}X_4 + a_{15}X_5 + a_{16}X_6 + a_{17}X_7$$
$$[X_2,U] = 0$$
$$[X_3,U] = a_{32}X_2$$
$$[X_4,U] = a_{42}X_2 + a_{32}X_3$$
$$[X_5,U] = a_{52}X_2 + a_{42}X_3 + a_{32}X_4$$
$$[X_6,U] = a_{62}X_2 + a_{52}X_3 + a_{42}X_4 + a_{32}X_5$$
$$[X_7,U] = a_{72}X_2 + a_{62}X_3 + a_{52}X_4 + a_{42}X_5 + a_{32}X_6$$
$$[X_8,U] = a_{82}X_2 + (a_{72} - a_{17} - a_{14})X_3 + (a_{62} - a_{15})X_4 + (a_{52} - a_{16})X_5 + (a_{42} - a_{17})X_6 + a_{32}X_7$$

(2.10) The FLA of dimension 8 with law $\mu_8^{18,\alpha}$ is not derived from any LA, since:

$$[X_1,U] = a_{12}X_2 + a_{13}X_3 + a_{14}X_4 + a_{15}X_5 + a_{16}X_6 + a_{17}X_7 + a_{18}X_8$$
$$[X_2,U] = 0$$
$$[X_3,U] = a_{32}X_2$$
$$[X_4,U] = a_{42}X_2 + a_{32}X_3$$
$$[X_5,U] = a_{52}X_2 + (a_{42} + \alpha\, a_{18})X_3 + a_{32}X_4$$
$$[X_6,U] = a_{62}X_2 + (a_{52} + a_{17} + a_{18})X_3 + (a_{42} + (1 + 2\alpha)\, a_{18})X_4 + a_{32}X_5$$
$$[X_7,U] = a_{72}X_2 + (a_{62} - a_{16})X_3 + (a_{52} + a_{17} + 2\,a_{18})X_4 + (a_{42} + (2+3\alpha)\, a_{18})X_5 + a_{32}X_6$$
$$[X_8,U] = a_{82}X_2 + (a_{72} - \alpha\, a_{15} - a_{16})X_3 + (a_{62} - (2+\alpha)\, a_{16} - a_{17})X_4 + (a_{52} + 2\,a_{18} - \alpha\, a_{17})X_5 + (a_{42} + (2+3\alpha)\, a_{18})X_6 + a_{32}X_7$$

NOTE: In this algebra it is verified that

$$a_{42} = 3/2\ (\alpha - 1)\ (\alpha - 2/3)$$

(2.11) The FLA of dimension 8 with law $\mu_8^{15,\alpha}$ is not derived from any LA, since:

$$[X_1,U] = a_{12}X_2 + a_{13}X_3 + a_{14}X_4 + a_{15}X_5 + a_{16}X_6 + a_{17}X_7$$
$$[X_2,U] = 0$$
$$[X_3,U] = a_{32}X_2$$
$$[X_4,U] = a_{42}X_2 + a_{32}X_3$$
$$[X_5,U] = a_{52}X_2 + a_{42}X_3 + a_{32}X_4$$
$$[X_6,U] = a_{62}X_2 + a_{52}X_3 + a_{42}X_4 + a_{32}X_5$$
$$[X_7,U] = a_{72}X_2 + a_{62}X_3 + a_{52}X_4 + a_{42}X_5 + a_{32}X_6$$
$$[X_8,U] = a_{82}X_2 + (a_{72} - \alpha\, a_{17} - a_{16} - a_{15})X_3 + (a_{62} - a_{17} - a_{16})X_4 + a_{52}X_5 + a_{42}X_6 + a_{32}X_7$$

(2.12) The FLA of dimension 8 with law μ_8^{17} is not derived from any LA, since:

$$[X_1,U] = a_{12} X_2 + a_{13} X_3 + a_{14} X_4 + a_{15} X_5 + a_{16} X_6 + a_{17} X_7 + a_{18} X_8$$
$$[X_2,U] = 0$$
$$[X_3,U] = a_{32} X_2$$
$$[X_4,U] = a_{42} X_2 + a_{32} X_3$$
$$[X_5,U] = a_{52} X_2 + a_{42} X_3 + a_{32} X_4$$
$$[X_6,U] = a_{62} X_2 + (a_{52} + a_{18}) X_3 + a_{42} X_4 + a_{32} X_5$$
$$[X_7,U] = a_{72} X_2 + (a_{62} + a_{18}) X_3 + (a_{52} + 2 a_{18}) X_4 + a_{42} X_5 + a_{32} X_6$$
$$[X_8,U] = a_{82} X_2 + (a_{72} - a_{17} - a_{16}) X_3 + (a_{62} + a_{18} - a_{17}) X_4 + (a_{52} + 2 a_{18}) X_5 + a_{42} X_6 + a_{32} X_7$$

■

Corollary 1.- There exist exactly 12 CNFLA of dimension 8. These are the following: μ_8^1, μ_8^2, μ_8^3, μ_8^4, $\mu_8^{6,\alpha}$, μ_8^8, $\mu_8^{9,\alpha}$, $\mu_8^{10,\alpha}$, μ_8^{11}, $\mu_8^{13,\alpha}$, $\mu_8^{15,\alpha}$ and μ_8^{17}.

Proof:

It is an immediate consequence of Theorems 1 and (1.3).

■

Theorem 2.- The complex FLA of dimension 8 with laws μ_8^5, $\mu_8^{7,\alpha}$, μ_8^{12}, $\mu_8^{14,\alpha}$, μ_8^{16}, μ_8^{18}, μ_8^{19} and μ_8^{20} are derived from SLA of dimension 9.

Proof:

By applying the method mentioned above to each of them we find that each of them is respectively derived from a SLA of dimension 9, with basis $\{X_1, \ldots, X_8, U\}$, verifying $[X_i, U] = a_i X_i$ $(1 \leq i \leq 8)$ where the coefficients a_i for each of them are the following:

FLA	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	obs.
μ_8^5	1	7	6	5	4	3	2	1	
$\mu_8^{7,\alpha}$	1	8	7	6	5	4	3	2	
μ_8^{12}	1	8	7	6	5	4	3	2	
$\mu_8^{14,\alpha}$	1	9	8	7	6	5	4	3	
μ_8^{16}	3	27	24	21	18	15	12	9	(*)
μ_8^{18}	1	10	9	8	7	6	5	4	
μ_8^{19}	1	11	10	9	8	7	6	5	
μ_8^{20}	1	8	7	6	5	4	3	2	

(*) In fact, if we apply the method mentioned above to this FLA we do not obtain these coefficients directly, but instead obtain the following expressions:

$$[X_1,U] = 3X_1 + 2X_8 \qquad [X_2,U] = 27X_2$$
$$[X_3,U] = 24X_3 \qquad [X_4,U] = 21X_4$$
$$[X_5,U] = 2X_3 + 18X_5 \qquad [X_6,U] = 4X_4 + 15X_6$$
$$[X_7,U] = 2X_3 + 6X_5 + 12X_7 \qquad [X_8,U] = 2X_4 + 6X_6 + 9X_8$$

Starting from these and by a suitable procedure we can obtain new fields Y_i, deduced from fields X_i, such that $[Y_i,U] = K_i\ Y_i$ (with $K_i \in C$); in fact, if we call respectively

$$Y_1 = 6X_1 - 2X_8 + X_6 \quad ; \qquad Y_5 = X_3 - 3X_5$$
$$Y_6 = 2X_4 - 3X_6 \quad ; \qquad Y_7 = X_5 - X_7$$
$$Y_8 = X_4 - 6X_6 + 6X_8$$

we obtain:

$$[Y_1,U] = 3Y_1 \qquad [Y_5,U] = 18Y_5 \qquad [Y_6,U] = 15Y_6$$
$$[Y_7,U] = 12Y_7 \qquad [Y_8,U] = 9Y_8$$

■

REFERENCES

[1] J. M. Ancochea y M. Goze. "Classification des algèbres de Lie filiformes de dimension 8".Arch.Math. vol.50. pp.511-525 (1988).

[2] J. M. Ancochea y M. Goze. "Classification des algèbres de Lie nilpotentes complexes de dim.7".Arch.Math.v.52.pp.175-185(1989).

[3] F. J. Echarte y J. Núñez. "Relations among solvable Lie Algebras and their nilpotent derived algebras". Proc.Sixth Int.Collo. on Dif. Geom. Univ.Santiago Compostela. pp.69-73.(1989).

[4] F. J. Echarte, J. R. Gómez Martín y J. Núñez. "Relación de Algebras de Lie Nilpotentes Complejas de dimensión 7 según sean o no derivadas de otras". Public. Sem. Mat. "García de Galdeano" Univ. Zaragoza. Sección I, n° 4 (1992). (Preprint).

[5] J. R. Gómez y F. J. Echarte. "Classification of complex Filiform Nilpotent Lie Algebras of dimension 9". Rendiconti Univ. Cagliari. vol. 61, Fasc. 1. (1991).

[6] Yu. B. Khakimdzhanov. "On Characteristically Nilpotent Lie Algebras". Soviet. Math. Dokl. vol.41 No. 1. pp.105-107. (1980).

[7] E. M. Luks. "A Characteristically Nilpotent Lie Algebra can be a derived Algebra". Proc. A.M.S. vol. 56 pp. 42-44. (1976).

[8] J. Núñez. "Las Algebras de Lie Filiformes complejas según sean o no derivadas de otras". Tesis Univ. Sevilla (1991).

[9] G. Valeiras. "Sobre las componentes irreducibles de la variedad de leyes de Algebras de Lie Nilpotentes complejas de dimensión 8 Tesis Univ. Sevilla (1992).

[10] M. Vergne. "Sur la varietè des lois nilpotentes". Bull. Soc. Math. France (1970).

APPENDIX

Algebras de Lie Filiformes Complejas de dimensión 8

μ_8^1 :
$$[X_1, X_i] = X_{i-1} \quad 3 \le i \le 8$$
$$[X_4, X_7] = X_2$$
$$[X_4, X_8] = X_2 + X_3$$
$$[X_5, X_6] = -X_2$$
$$[X_5, X_7] = -(2/5)\, X_2$$
$$[X_5, X_8] = X_4 + (3/5)\, X_3$$
$$[X_6, X_7] = -(2/5)\, X_3$$
$$[X_6, X_8] = X_5 + (1/5)\, X_4$$
$$[X_7, X_8] = X_6 + (1/5)\, X_5$$

μ_8^2 :
$$[X_1, X_i] = X_{i-1} \quad 3 \le i \le 8$$
$$[X_4, X_7] = X_2$$
$$[X_4, X_8] = X_3$$
$$[X_5, X_6] = -X_2$$
$$[X_5, X_8] = X_4$$
$$[X_6, X_7] = X_2$$
$$[X_6, X_8] = X_2 + X_3 + X_5$$
$$[X_7, X_8] = X_3 + X_4 + X_6$$

μ_8^3 :
$$[X_1, X_i] = X_{i-1} \quad 3 \le i \le 8$$
$$[X_4, X_7] = X_2$$
$$[X_4, X_8] = X_3$$
$$[X_5, X_6] = -X_2$$
$$[X_5, X_8] = X_4$$
$$[X_6, X_8] = X_2 + X_5$$
$$[X_7, X_8] = X_3 + X_6$$

μ_8^4 : $[X_1, X_i] = X_{i-1}$ $\quad 3 \le i \le 8$
$[X_4, X_7] = X_2$
$[X_4, X_8] = X_3$
$[X_5, X_6] = -X_2$
$[X_5, X_8] = X_4$
$[X_6, X_7] = X_2$
$[X_6, X_8] = X_3 + X_5$
$[X_7, X_8] = X_4 + X_6$

μ_8^5 : $[X_1, X_i] = X_{i-1}$ $\quad 3 \le i \le 8$
$[X_4, X_7] = X_2$
$[X_5, X_6] = -X_2$
$[X_i, X_8] = X_{i-1}$ $\quad 4 \le i \le 7$

$\mu_8^{6,\alpha}$: $[X_1, X_i] = X_{i-1}$ $\quad 3 \le i \le 8$
$[X_4, X_8] = \alpha X_2$
$[X_5, X_7] = X_2$ $\quad \alpha \in C-[-1]$
$[X_5, X_8] = (1+\alpha) X_3 + X_2$
$[X_6, X_7] = X_3$
$[X_6, X_8] = (2+\alpha) X_4 + X_3$
$[X_7, X_8] = (2+\alpha) X_5 + X_4$

$\mu_8^{7,\alpha}$: $[X_1, X_i] = X_{i-1}$ $\quad 3 \le i \le 8$
$[X_4, X_8] = \alpha X_2$
$[X_5, X_7] = X_2$
$[X_5, X_8] = (1+\alpha) X_3$ $\quad \alpha \in C$
$[X_6, X_7] = X_3$
$[X_6, X_8] = (2+\alpha) X_4$
$[X_7, X_8] = (2+\alpha) X_5$

μ_8^8 : $[X_1, X_i] = X_{i-1}$ $\quad 3 \le i \le 8$
$[X_4, X_8] = -X_2$
$[X_5, X_7] = X_2$
$[X_6, X_7] = X_2 + X_3$
$[X_6, X_8] = X_3 + X_4$
$[X_7, X_8] = X_4 + X_5$

$\mu_8^{9,\alpha}$: $[X_1, X_i] = X_{i-1}$ $\quad 3 \le i \le 8$
$[X_4, X_8] = X_2$
$[X_5, X_8] = X_3$
$[X_6, X_7] = X_2$ $\quad \alpha \in C$
$[X_6, X_8] = \alpha X_2 + X_3 + X_4$
$[X_7, X_8] = \alpha X_3 + X_4 + X_5$

$$\mu_8^{10,\alpha}: [X_1, X_i] = X_{i-1} \quad 3 \le i \le 8$$
$$[X_4, X_8] = X_2$$
$$[X_5, X_8] = X_3 \qquad \alpha \in C$$
$$[X_6, X_8] = X_2 + X_4$$
$$[X_7, X_8] = \alpha X_2 + X_3 + X_5$$

$$\mu_8^{11}: [X_1, X_i] = X_{i-1} \quad 3 \le i \le 8$$
$$[X_i, X_8] = X_{i-2} \quad 4 \le i \le 6$$
$$[X_7, X_8] = X_2 + X_5$$

$$\mu_8^{12}: [X_1, X_i] = X_{i-1} \quad 3 \le i \le 8$$
$$[X_i, X_8] = X_{i-1} \quad 4 \le i \le 7$$

$$\mu_8^{13,\alpha}: [X_1, X_i] = X_{i-1} \quad 3 \le i \le 8$$
$$[X_5, X_8] = \alpha X_2$$
$$[X_6, X_7] = X_2 \qquad \alpha \in C$$
$$[X_6, X_8] = (1+\alpha) X_3$$
$$[X_7, X_8] = (1+\alpha) X_4 + X_3$$

$$\mu_8^{14,\alpha}: [X_1, X_i] = X_{i-1} \quad 3 \le i \le 8$$
$$[X_5, X_8] = \alpha X_2$$
$$[X_6, X_7] = X_2 \qquad \alpha \in C$$
$$[X_6, X_8] = (1+\alpha) X_3$$
$$[X_7, X_8] = (1+\alpha) X_4$$

$$\mu_8^{15,\alpha}: [X_1, X_i] = X_{i-1} \quad 3 \le i \le 8$$
$$[X_5, X_8] = X_2$$
$$[X_6, X_8] = X_2 + X_3 \qquad \alpha \in C$$
$$[X_7, X_8] = \alpha X_2 + X_3 + X_4$$

$$\mu_8^{16}: [X_1, X_i] = X_{i-1} \quad 3 \le i \le 8$$
$$[X_5, X_8] = X_2$$
$$[X_6, X_8] = X_3$$
$$[X_7, X_8] = X_2 + X_4$$

$$\mu_8^{17}: [X_1, X_i] = X_{i-1} \quad 3 \le i \le 8$$
$$[X_6, X_8] = X_2$$
$$[X_7, X_8] = X_2 + X_3$$

$$\mu_8^{18}: [X_1, X_i] = X_{i-1} \quad 3 \le i \le 8$$
$$[X_6, X_8] = X_2$$
$$[X_7, X_8] = X_3$$

$$\mu_8^{19}: [X_1, X_i] = X_{i-1} \quad 3 \le i \le 8$$
$$[X_7, X_8] = X_2$$

$$\mu_8^{20}: [X_1, X_i] = X_{i-1} \quad 3 \le i \le 8$$

Rev. Academia de Ciencias. Zaragoza. 48 (1993)

On the Holonomy Bundle of the Sphere, II

by

Fernando Etayo and Ujué R. Trías

Departamento de Matemáticas, Estadística y Computación

Facultad de Ciencias. Universidad de Cantabria

Avda. de los Castros, s/n. 39071 SANTANDER

ABSTRACT. We shall prove that the holonomy bundle of the sphere S^n is $SO(n+1) \longrightarrow S^n$ with fibre $SO(n)$.

A.M.S. Math. Subj. Class. 53C05.

§1.Introduction.

The holonomy bundle of the sphere S^2, with the canonical metric, is the real projective 3-space, with fibre S^1 (see [1], [2], [4] for different proofs). By using the quotient manifold structure of the sphere S^n we shall obtain its holonomy bundle.

We sall use the following notations (see [1]): S^n is the unit sphere immersed in $\mathbb{R}^{n+1}$, with the canonical metric. As homogeneous space, S^n is the quotient manifold $SO(n+1)/SO(n)$. Moreover, $F(S^n) \longrightarrow S^n$ and $SO(S^n) \longrightarrow S^n$ are the frame bundle and the bundle of positively oriented orthonormal frames. If $u \in F(S^n)$, then $\mathcal{H}_u \longrightarrow S^n$ is the holonomy bundle at u, which is a principal bundle with fibre the holonomy group $SO(n)$ ([3], [5]).

§2. The result.

LEMMA. $SO(S^n) \longrightarrow S^n$ *is a principal bundle isomorphic to the principal bundle* $SO(n+1) \longrightarrow S^n$.

Proof. An $(n+1)\times(n+1)$ square matrix $A \in SO(n+1)$ can be considered as an orthonormal positively oriented basis $\{p, v_1, \ldots, v_n\}$ of $\mathbb{R}^{n+1}$, when it is given by columnes. Then, $p \in S^n$ and $\{v_1, \ldots, v_n\}$ is an orthonormal positive oriented basis of the tangent space $T_p S^n$. The bundle isomorphism is given by α: $SO(n+1) \longrightarrow SO(S^n)$, $\alpha(A)=\{v_1, \ldots, v_n\}$.

QED

THEOREM. *The holonomy bundle* $\mathcal{H}_u \longrightarrow S^n$ *is isomorphic to the principal bundle* $SO(n+1) \longrightarrow S^n$.

Proof. Let $u \in SO(S^n) \subset F(S^n)$. We shall prove that $\mathcal{H}_u = SO(S^n)$. Let us assume that u is an orthonormal and positively oriented basis of T_pS^n, and let u' an orthonormal and positively oriented basis of $T_{p'}S^n$. Let γ be a path joining p and p'. By parallel transport along γ, u is moved to a new orthonormal and positively oriented basis w of $T_{p'}S^n$. Then, there exists a closed path η on p' such that w is moved to u' by parallel transport along η. Then, $SO(S^n) \subset \mathcal{H}_u$. The other inclussion is well known.

If u is not in $SO(S^n)$, the result follows from the following property (see [3], [5]): All of the holonomy bundles are isomorphic.

QED

Examples.

n=2: $S^2=SO(3)/SO(2)$; $\mathcal{H}_u \cong SO(3) \cong P_3(\mathbb{R})$, as it is known ([1], [2], [4]).

n=3: $S^3=SO(4)/SO(3)$; $\mathcal{H}_u \cong SO(4) \cong S^3 \times SO(3) \cong S^3 \times P_3(\mathbb{R})$ is a trivial bundle.

n=7: $S^7=SO(8)/SO(7)$; $\mathcal{H}_u \cong SO(8) \cong S^7 \times SO(7)$ is a trivial bundle.

References.

[1] Etayo, F.; Trías, U. R.: On the Holonomy Bundle of the Sphere. *Rev. Acad. Ciencias Zaragoza* **47** (1992) 83–87.

[2] Klingerberg, W.; Sasaki, S.: On the Tangent Sphere Bundle of a 2-Sphere. *Tôhoku Math. Journ.***27** (1975) 49–56.

[3] Kobayashi, S.; Nomizu, K.: *Foundations of Differential Geometry I.* J. Wiley, N. York, 1963.

[4] Montesinos, J. M.: *Classical Tesellations and Three-Manifolds.* Springer, 1985.

[5] Poor, W. A.: *Differential Geometric Structures.* McGraw Hill, N. York, 1981.

Rev. Academia de Ciencias. Zaragoza. 48 (1993)

ON WARING'S PROBLEM *

BY

C. CALDERÓN AND M.J. DE VELASCO

Universidad del País Vasco. Facultad de Ciencias. Departamento de Matemáticas.
E-48080 Bilbao - España .

Fecha de entrega : 19-12-92

Abstract . The object of present paper is to study the Waring's problem concerning solvability of the equation $X_1^s+\ldots+X_r^s = N$ in nonnegative integers $X_1,\ldots,X_r$ for every sufficiently large natural number N if $r \geq r(s)$ and $X_1,\ldots,X_r$ are on a determinated arithmetic progression .

1. INTRODUCTION AND NOTATION .

Among the most famous problems over representing a number as a sum of other numbers is Waring's Conjecture: *Every integer $N > 0$ is the sum of a fixed least number $g(s)$ of powers of integers that are greater than or equal to zero* . Loo Keng Hua constructed in [4] a generalization for polynomials of the Hardy-Littlewood asymptotic formula in Waring's Conjecture. Thus, he studied the number of solutions of the Diophantine equation $N = P_1(h_1) + \ldots + P_r(h_r)$, $(h_\nu \geq 0)$ where P's are integral-valued polynomials of the k-th degree, with positive highest coefficient. In [8] Mit'kin obtained a sharp upper bound for the smallest r for which the equation $N = f(x_1) + \ldots + f(x_r)$ is solvable in nonnegative integers $x_1,\ldots,x_r$ where $f(x) = a_n\binom{x}{n} + \ldots + a_1\binom{x}{1}$ and $(a_n,\ldots,a_1) = 1$. The circle method due to Hardy , Littlewood and Ramanujan with Vinogradov trigonometric sums , is a powerful tool in the study of Diophantine problems of the type $f(x_1,\ldots,x_r) = 0$ when the number of variables is large compared with the degree of the polynomial f. In this paper we wish to find an asymptotic formula for the number of solutions of the Diophantine equation

$$X_1^s + \ldots + X_r^s = N, \text{with} \quad 0 \leq X_i \leq P, \quad 1 \leq i \leq r, \quad X_i \equiv z_i (\mathrm{mod}\, Q) \tag{1.1}$$

where each $z_j, 1 \leq j \leq r$ runs through a complete set of residues mod Q , and $(N, Q) = 1$

* Supported by the University of the Basque Country

Let us denote by $R(N)$ the number of solutions of (1.1) . For convenience we introduce the notation $e^{2\pi it} = e(t)$. Consider the function

$$f_j(\rho) = \sum_{\substack{X_j=1 \\ X_j \equiv z_j \pmod{Q}}}^{P} e(\rho X_j^s), 1 \le i \le r \tag{1.2}$$

where ρ is a real number . By elementary calculus it is easy to find an expression for $R(N)$, thus

$$R(N) = \int_0^1 \left(\prod_{j=1}^{r} f_j(\rho) \right) e(-\rho N) d\rho. \tag{1.3}$$

The idea of the method is to show that if r , depending only of s , is chosen sufficiently large , then the integral (1.3) will be is positive for all $N \ge N_0(s)$. This means that the equation (1.1) has at least one solution for all $N \ge N_0(s)$. We notice that the integrand is a function with period 1 and so the value of the integral is the same if we integrate over any interval of length 1. If $\rho \epsilon [-1/\tau, 1 - 1/\tau)$, from Dirichlet's Lemma ([6]), we can write, for $\tau = P^{s-\frac{1}{2}} Q^r$:

$$\rho = \frac{a}{q} + z, \quad (a,q) = 1 \quad ; 1 \le q \le \tau; \quad |z| \le \frac{1}{q\tau}. \tag{1.4}$$

Let $\mathfrak{M}(a,q) = \{\rho;\ |\rho - \frac{a}{q}| < \frac{1}{q\tau}\}$. The set $\mathfrak{M} = \cup \mathfrak{M}(a,q)$ such that $1 \le q \le P^{\frac{1}{8}}$ is called major arc . The minor arc is its complement $\mathfrak{m} = [0,1] \setminus \mathfrak{M}$ and holds $P^{\frac{1}{8}} < q \le \tau$. Then ,

$$R(N) = R_{\mathfrak{M}}(N) + R_{\mathfrak{m}}(N). \tag{1.5}$$

To show that $R(N)$ is positive if r is enough large , we shall obtain an asymptotic formula for $R(N)$

$$R(N) = \frac{(\Gamma(1+\frac{1}{s}))^r}{\Gamma(r/s)} \frac{P^{r-s}}{Q^r} G(N) + O(\frac{P^{r-s-\frac{1}{2^{s+4}}}}{Q^r}) \tag{1.6}$$

provided $r \ge c_3 s^2 \ln s$, $s > 2$ and $Q \le P^c$ with $c \le \min\{\frac{1}{4r}, \frac{1}{2^{s+4}}\}$. Here , if we take $P = [N^{\frac{1}{s}}]$, and if N is sufficiently large then the first term is always larger than the error term , so $R(N) > 0$.

2. AN ASYMPTOTIC FORMULA FOR R(N)

In each interval $\mathfrak{M}(a,q)$ we find a function which approximates to $\prod_{j=1}^{r} f_j(\rho)$. This is given by the following lemma .

Lemma 1. *If* $\rho \epsilon \mathfrak{M}(a,q)$, *then*

$$\prod_{j=1}^{r} f_j(\rho) = \frac{I_z^r}{Q^r} \prod_{j=1}^{r} S_j(a,q;z_j) + \theta_1 \left(\frac{P^{r-1}}{(Q_1 \dots Q_r)^{1/s}} qQ\right), \quad |\theta_1| \leq 1$$

where

$$S_j(a,q;z_j) = \frac{(Q,q)}{q} \sum_{\substack{x_j=1 \\ x_j \equiv z_j \,(mod\,(Q,q))}}^{q} e\left(\frac{a}{q} x_j^s\right). \quad j = 1, \dots, r$$

and $I_z = \int_0^P e(z\xi^s)d\xi$.

Proof . The equality

$$\prod_{j=1}^{r} f_j(\rho) = \frac{1}{Q^r} \prod_{j=1}^{r} \sum_{X_j=1}^{P} \sum_{d_j=1}^{Q} e\left(\rho X_j^s + \frac{d_j(X_j - z_j)}{Q}\right) \tag{2.1}$$

is deduced easily from the ortogonality relations

$$\frac{1}{Q} \sum_{d=1}^{Q} e\left(\frac{\ell d}{Q}\right) = \begin{cases} 1, & \text{if } \ell \equiv 0 \pmod{Q} \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

Each $d_j, 1 \leq j \leq r$ runs through a complete set of residues mod Q . We denote $\delta_j = g.c.d.(d_j, Q) = (d_j, Q)$ and $Q_j = \ell.c.m.(Q/\delta_j, q) = [Q/\delta_j, q]$. We can make the following change of the variables of summation in (2.1) : $X_j = x_j + Q_j a_j$, $1 \leq x_j \leq Q_j$, $-x_j Q_j^{-1} < a_j \leq (P - x_j)Q_j^{-1}$, $1 \leq j \leq r$. Then

$$\prod_{j=1}^{r} f_j(\rho) = \frac{1}{Q^r} \prod_{j=1}^{r} \sum_{d_j=1}^{Q} \sum_{x_j=1}^{Q_j} e\left(\frac{a}{q} x_j^s + \frac{d_j(x_j - z_j)}{Q}\right) \sum_{a_j} e(z(Q_j a_j + x_j)^s)dz \tag{2.3}$$

Moreover

$$\left|\frac{\partial}{\partial a_j}(z((Q_1 a_1 + x_1)^s + \dots + (Q_r a_r + x_r)^s))\right| \leq |z| s Q_j P^{s-1}$$

$$\leq \frac{1}{\tau} s \frac{Q}{\delta_j} P^{s-1} \leq \frac{sQ}{\tau} P^{s-1} < 0'5 \quad j = 1, \dots, r.$$

Hence from the generalization of the van der Corput Lemma by Arhipov-Karatzuba-Chubarikov (Lemma -3 [2]), we have

$$\prod_{j=1}^{r} \sum_{-x_j Q_j^{-1} < a_j \leq (P-x_j)Q_j^{-1}} e(z(Q_j a_j + x_j)^s) =$$

$$= \int_{\frac{-x_1}{Q_1}}^{\frac{P-x_1}{Q_1}} \cdots \int_{\frac{-x_r}{Q_r}}^{\frac{P-x_r}{Q_r}} e(z((x_1+Q_1a_1)^s+\ldots+(x_r+Q_ra_r)^s))da_1 \ldots da_r + 2\theta_2 rs \frac{P^{r-1}}{Q_1 \ldots Q_r} qQ$$

$$(2.4) \qquad = \frac{1}{Q_1 \ldots Q_r} I_z^r + 2\theta_2 rs \frac{P^{r-1}}{Q_1 \ldots Q_r} qQ, \quad I_z = \int_0^P e(z\xi^s)d\xi, \quad |\theta_2| \leq 1.$$

In 1985, Minggao-Ding Ping proved in [7] that :

$$\sum_{x=1}^{q} e\left(\frac{f(x)}{q}\right) \leq e^{2s} q^{1-\frac{1}{s}}$$

where $f(x)$ is an integer- valued polynomial of degree s. Then we can apply this estimate for the trigonometric rational sum :

$$(2.5) \qquad \sum_{x_j=1}^{Q_j} e\left(\frac{F_j(x_j)}{Q_j}\right) \leq e^{2s} Q_j^{1-1/s}$$

where

$$(2.6) \qquad \frac{F_j(x_j)}{Q_j} = \frac{a}{q}x_j^s + \frac{d_j x_j}{Q} = \frac{a'_j x_j^s + b'_j x_j}{Q_j}, (a'_j, b'_j, Q_j) = 1, \quad j = 1, \ldots, r$$

In consequence replacing (2.4) , (2.6) in (2.3) we deduce

$$\prod_{j=1}^{r} f_j(\rho) = \frac{I_z^r}{Q^r} \prod_{j=1}^{r} \sum_{d_j=1}^{Q} \frac{1}{Q_j} \sum_{x_j=1}^{Q_j} e(\frac{a}{q}x_j + \frac{d_j(x_j - z_j)}{Q}) + \theta_1 (\frac{P^{r-1}}{(Q_1 \ldots Q_r)^{1/s}} qQ),$$

where $|\theta_1| \leq 1$. If $\delta_j = g.c.d.(d_j, Q) = (d_j, Q)$ then $Q = Q'\delta_j$, $d_j = d'_j\delta_j$ with $(Q', d'_j) = 1$, $j = 1, \ldots, r$ and by the properties of the arithmetical functions we deduce

$$\sum_{d_j=1}^{Q} \frac{1}{Q_j} \sum_{x_j=1}^{Q_j} e(\frac{a}{q}x_j + \frac{d_j(x_j - z_j)}{Q}) = \sum_{\delta_j | Q} \frac{1}{[q, \delta_j]} \sum_{x_j=1}^{[q,\delta_j]} e(\frac{a}{q}x_j^s) \sum_{\substack{d'_j=1 \\ (d'_j, \delta_j)=1}}^{\delta_j} e(\frac{(x_j - z_j)d'_j}{\delta_j}) =$$

$$= \sum_{\delta_j | Q} \frac{1}{[q, \delta_j]} \sum_{x_j=1}^{[q,\delta_j]} c_{\delta_j}(x_j - z_j) e(\frac{a}{q}x_j^s) = S_j(a, q; z_j)$$

where $c_\delta(x-z)$ denotes the Ramanujan sum . We make the following change of variable $x_j = uq + v$; $\quad 1 \leq v \leq q, \quad 1 \leq u \leq \frac{[q,\delta_j]}{q}$. Therefore we obtain

$$S_j(a,q;z_j)=\frac{(q,Q)}{q}\sum_{v=1}^{q}e(\frac{a}{q}v^s);\quad \text{if}\quad (q,Q)|v-z_j$$

and $S_j(a,q;z_j)=0$ otherwise. Thus, the sum $S_j(a,q;z_j)$ holds

$$S_j(a,q;z_j)=\frac{(Q,q)}{q}\sum_{\substack{x_j=1\\ x_j\equiv z_j(\mathrm{mod}\,(Q,q))}}^{q}e(\frac{a}{q}x_j^s).$$

The lemma is proved . ∎

Lemma 2. *For $r>2s$, and $S_j(a,q;z_j)$ given in the above Lemma the series*

$$G(z_1,\dots,z_r;N)=\sum_{q=1}^{\infty}\sum_{\substack{a\le q\\(a,q)=1}}e\left(-\frac{a}{q}N\right)\left(\prod_{j=1}^{r}S_j(a,q;z_j)\right) \tag{2.7}$$

is absolutely convergent .

Proof. From Lemma 1 we have that $G(z_1,\dots,z_r;N)$ satisfies the following equality

$$G(z_1,\dots,z_r;N)=\sum_{q=1}^{\infty}\sum_{\substack{a\le q\\(a,q)=1}}e(-\frac{a}{q}N)\prod_{j=1}^{r}\frac{(Q,q)}{q}\sum_{\substack{x_j=1\\ x_j\equiv z_j(\mathrm{mod}\,(Q,q))}}^{q}e(\frac{a}{q}x_j^s).$$

Moreover , from (2.2), we can write

$$\sum_{\substack{x=1\\ x\equiv z(\mathrm{mod}\,(q,Q))}}^{q}e\left(\frac{a}{q}x^s\right)=\sum_{x=1}^{q}e(\frac{a}{q}x^s)\frac{1}{(q,Q)}\sum_{h=1}^{(q,Q)}e\left(h\frac{(x-z)}{(q,Q)}\right)$$

From (2.5) we have

$$|\sum_{\substack{a\le q\\(a,q)=1}}e\left(-\frac{a}{q}N\right)\left(\prod_{j=1}^{r}S_j(a,q;z_j)\right)|\ll$$

$$\ll\sum_{\substack{a\le q\\(a,q)=1}}\prod_{j=1}^{r}\frac{1}{q}\sum_{h=1}^{(q,Q)}q^{1-1/s}\ll Q^rq^{-r/s+1}.$$

And the lemma is proved. ∎

The next lemma is related to the major arcs .

Lemma 3. *Let $r\ge c_1s^2\ln s$, $s>2$ and $Q\le P^c$ with $c\le\min\{\frac{1}{4r},\frac{1}{2^{s+4}}\}$, $P=[N^{\frac{1}{s}}]$. Then we have the following asymptotic formula uniform on Q.*

$$(2.8)\qquad R_{\mathfrak{M}}(N)=\frac{\left(\Gamma(1+\frac{1}{s})\right)^r}{\Gamma(r/s)}\frac{P^{r-s}}{Q^r}G(z_1,\ldots,z_r;N)+O(\frac{P^{r-s-1/4(1-1/r)}}{Q^r})$$

Proof . From (1.3) , (1.4) and Lemmas 1-2 we have:

$$(2.9)\qquad R_{\mathfrak{M}}(N)=\int_{\mathfrak{M}}\left(\prod_{j=1}^{r}f_j(\rho)\right)e(-\rho N)d\rho=$$

$$=\frac{1}{Q^r}\sum_{q\le P^{1/8}}\sum_{\substack{a\le q\\(a,q)=1}}e(-\frac{a}{q}N)\left(\prod_{j=1}^{r}S_j(a,q;z_j)\right)\int_{-\frac{1}{q\tau}}^{\frac{1}{q\tau}}e(-zN)I_z^r dz+$$

$$+O\left(\sum_{q\le P^{1/8}}\sum_{\substack{a\le q\\(a,q)=1}}\frac{P^{r-1}}{q\tau}qQ\right).$$

The O-term holds

$$(2.10)\qquad \sum_{q\le P^{1/8}}\sum_{\substack{a\le q\\(a,q)=1}}\frac{P^{r-1}}{q\tau}qQ\ll\frac{P^{r-1}}{\tau}\sum_{q\le P^{1/8}}qQ\ll\frac{P^{r-s-1/4(1-1/r)}}{Q^r}.$$

I.M. Vinogradov (Lemma - 4, Chapter 2, [9]) proved that :

$$\left|\int_0^1 e(f(x))\,dx\right|\le min\{1,62nu^{\frac{1}{n}}\}$$

where $f(x)=u_nx^n+\ldots+u_1x$ and $u=max\{u_i/\ 1\le i\le n\}$. Using this estimate, we obtain:

$$\int_{\frac{1}{q\tau}}^{\infty}I_z^r dz\ll\int_{1/q\tau}^{\infty}z^{-\frac{r}{s}}dz\ll(q\tau)^{\frac{r}{s}-1}\ll q^{r/s-1}P^{r-s-1/4(r/s-1)}$$

Moreover :

$$\int_{-\infty}^{+\infty}e(-zN)I_z^r=\frac{\left(\Gamma(1+\frac{1}{s})\right)^r}{\Gamma(r/s)}P^{r-s}+O(P^{r-s-1})$$

(see Th-6.3, Ch - 2, Ayoub [3]). Hence:

$$(2.11)\qquad \int_{-\frac{1}{q\tau}}^{\frac{1}{q\tau}}e(-zN)I_z^r dz=\frac{\left(\Gamma(1+\frac{1}{s})\right)^r}{\Gamma(r/s)}P^{r-s}+O(P^{r-s-1})+O(q^{r/s-1}P^{r-s-1/4(r/s-1)})$$

Now we go to study the sum

$$\sum_{q>P^{1/8}}\sum_{\substack{a\le q\\(a,q)=1}}e\left(-\frac{a}{q}N\right)\prod_{j=1}^{r}\frac{1}{Q_j}\sum_{x_j=1}^{Q_j}e\left(\frac{F_j(x_j)}{Q_j}\right).$$

From (2.5)

$$(2.12)\qquad \sum_{q>P^{1/8}} \sum_{\substack{a\le q\\(a,q)=1}} \prod_{j=1}^{r} \left| \frac{1}{Q_j} \sum_{x_j=1}^{Q_j} e\left(\frac{F_j(x_j)}{Q_j}\right) \right| \ll \frac{P^{1/2-r/8s}}{Q^r}$$

and sustituying (2.10) (2.11) and (2.12) in (2.9) and using Lemma 2 we deduce that

$$(2.13)\qquad R_{\mathfrak{M}} = \frac{\left(\Gamma(1+\frac{1}{s})\right)^r}{\Gamma(r/s)} \frac{P^{r-s}}{Q^r} G(z_1,\dots,z_r;N)$$

$$+O\left(\frac{P^{r-s-1/4(r/s-5/2)}}{Q^r}\right)+O\left(\frac{P^{r-s-1/4(1-1/r)}}{Q^r}\right)+O\left(\frac{P^{r-s-1/8(r/s-4)}}{Q^r}\right)$$

And under the hypothesis of the lemma from (2.13) we deduce (2.8) . ∎

The contribution of the minor arcs $\mathfrak{m}$ is

$$(2.14)\qquad R_{\mathfrak{m}}(N) = \int_{\mathfrak{m}} e(-\rho N) \prod_{j=1}^{r} \sum_{\substack{X_j=1\\ X_j\equiv z_j \,(\mathrm{mod}\, Q)}}^{P} e(\rho X_j^s) d\rho$$

which holds the following result :

Lemma 4. *If* $r \ge c_2 s^2 \ln s$, $s>2$, *then*

$$(2.15)\qquad R_{\mathfrak{m}}(N) \ll \frac{P^{r-s-\frac{1}{2^s+4}}}{Q^r}$$

Proof . From (2.14) we have

$$|R_{\mathfrak{m}}(N)| \le \int_{\mathfrak{m}} \prod_{j=1}^{r} \left| \sum_{\substack{X_j=1\\ X_j\equiv z_j \,(\mathrm{mod}\, Q)}}^{P} e(\rho X_j^s) \right| d\rho \le$$

$$\le \prod_{i=1}^{r-2k_1} \max_{\rho\in\mathfrak{m}} \Big| \sum_{\substack{X_i=1\\ X_i\equiv z_i \,(\mathrm{mod}\, Q)}}^{P} e(\rho X_i^s) \Big| \int_0^1 \prod_{j=1}^{2k_1} \Big| \sum_{\substack{X_j=1\\ X_j\equiv z_j \,(\mathrm{mod}\, Q)}}^{P} e(\rho X_j^s) \Big| d\rho$$

with $j = i + 2k_1 - r$. From Hölder inequality we obtain :

$$\le \prod_{i=1}^{r-2k_1} \max_{\rho\in\mathfrak{m}} \left| \sum_{\substack{X_i=1\\ X_i\equiv z_i \,(\mathrm{mod}\, Q)}}^{P} e(\rho X_i^s) \right| \prod_{j=1}^{2k_1} \left[\int_0^1 \left| \sum_{\substack{X_j=1\\ X_j\equiv z_j \,(\mathrm{mod}\, Q)}}^{P} e(\rho X_j^s) \right|^{2k_1} d\rho \right]^{1/2k_1}$$

Moreover $P^{1/8} < q < P^{s-1/8}$, and from Weyl inequality ([4]) we deduce that

$$\left| \sum_{\substack{X_i=1 \\ X_i \equiv z_i (\mathrm{mod}\, Q)}}^{P} e(\rho X_i^s) \right| \leq \frac{1}{Q} \left| \sum_{d=1}^{Q} \sum_{X_i=1}^{P} e(\rho X_i^s + \frac{dX_i}{Q}) \right| \ll P^{1-\frac{1}{2^s+3}}.$$

For the integral we obtain the following inequality

$$\int_0^1 | \sum_{\substack{X_j=1 \\ X_j \equiv z_j (\mathrm{mod}\, Q)}}^{P} e(\rho X_j^s)|^{2k_1} d\rho \leq \tag{2.16}$$

$$\leq \frac{1}{Q} \sum_{d=1}^{Q} \int_0^1 \sum_{\substack{Y_1,\ldots,Y_{k_1} \\ X_1,\ldots,X_{k_1}}} e(\rho(\sum_{i=1}^{k_1}(X_i^s - Y_i^s)) + \frac{d}{Q}(\sum_{i=1}^{k_1}(X_i - Y_i)))d\rho$$

Let T the number of solutions of the system

$$\left.\begin{aligned} X_1^s + \ldots + X_{k_1}^s - Y_1^s - \ldots - Y_{k_1}^s = 0 \\ X_1 + \ldots + X_{k_1} - Y_1 - \ldots - Y_{k_1} = hQ \end{aligned}\right\}$$

whose variables take the values

$1 \leq X_1, \ldots, X_{k_1}; Y_1 \ldots, Y_{k_1} \leq P$; $-k_1 P/Q < h < k_1 P/Q$

and we denote $J_{k_1,s}(\lambda_1, \ldots, \lambda_s)$ the number of solutions of the system

$$X_1^\nu + \ldots + X_{k_1}^\nu - Y_1^\nu - \ldots - Y_{k_1}^\nu = \lambda_\nu, \quad 1 \leq \nu \leq s$$

where $\lambda_1, \ldots, \lambda_s$ hold $|\lambda_\nu| < k_1 P^\nu; \nu = 1, \ldots, s$. We can write

$$T = \sum_h \sum_{\substack{\lambda_2,\ldots,\lambda_{s-1} \\ |\lambda_j| < k_1 P^j}} J_{k_1,s}(hQ, \lambda_2, \ldots, \lambda_{s-1}, 0) \leq (2k_1)^{s-1} \frac{P^{s(s-1)/2}}{Q} J_{k_1,s}(P)$$

where $J_{k_1,s}(P) = J_{k_1,s}(0, \ldots, 0) \leq e^{cs^3 \ln s} P^{2k_1 - s(s+1)/2}$ being $k_1 \geq c_0 s^2 \log s$, with c and c_0 are positive constants, (Arhipov-Karatzuba-Chubarikov [1]). Then

$$T \ll \frac{P^{2k_1-s}}{Q} \tag{2.17}$$

Therefore, from (2.16),(2.15) we deduce

$$|R_{\mathfrak{m}}(N)| \ll P^{(1-\frac{1}{2^s+3})(r-2k_1)} \frac{P^{2k_1-s}}{Q} \ll \frac{P^{r-s-\frac{1}{2^s+4}}}{Q^r} \tag{2.18}$$

and Lemma 4 is proved . ∎

In order to study $G(z_1,\ldots,z_r;N)$ it is necessary to prove the following lemmas concerning congruences and singular series .

Lemma 5. *If the equation* $x^s \equiv a(mod p^\beta)$ *with* $(p,s)=1$, $(x,p)=1$ *is soluble , then* $x^s \equiv a(mod p^m)$ $\forall m > \beta$ *is solvable also .*

The proof can be found in A.A.Karatzuba [6] (Chapter XI, pag 231)

Lemma 6. *Let* $T_\beta(p^m)$ *be the number of solutions of the congruence*

$$x_1^s + \ldots + x_r^s \equiv N(mod p^m)$$

with $x_j \equiv z_j(mod p^\beta)$; $\quad j = 1,\ldots,r; 1 \le \beta < m, \quad (p,N) = 1$. *Then*

$$T_\beta(p^m) = p^{(r-1)(m-\beta)}.$$

Proof . We can write each x_i in the form

$$x_i = z_i + x_{i,0}p^\beta + \ldots + x_{i,m-\beta-1}p^{m-1}; \quad 1 \le x_{i,j} \le p$$

Since $z_1^s + \ldots + z_r^s \equiv N(\mathrm{mod}\, p^\beta)$, we have

$$(z_1 + p^\beta x_{10})^s + \ldots + (z_r + p^\beta x_{r0})^s \equiv N(\mathrm{mod}\, p^{\beta+1}).$$

Thus , we can suppose that $p \nmid z_1$ and we obtain

$$x_{1,0} \equiv (z_1^{s-1})^{-1}(N_1^* - (z_2^{s-1}x_{2,0} + \ldots + z_r^{s-1}x_{r,0}))(\mathrm{mod}\, p)$$

For each r-1 different values of the variables $x_{2,0},\ldots,x_{r,0}$ there is only one value of $x_{1,0}$, that's p^{r-1} -tuples $(x_{1,0},\ldots,x_{r,0})$. Repeating this argument in an iterative form, the number of solutions holds

$$T_\beta(p^m) \le p^{(m-\beta)(r-1)}. \tag{2.19}$$

Now, we prove that $T_\beta(p^m) \ge p^{(m-\beta)(r-1)}$. In fact , as $z_1^s + \ldots + z_r^s \equiv N(\mathrm{mod}\, p^\beta)$ and $(p,N) = 1$ we have

$$x_1^s + (z_2 + p^\beta y_2)^s + \ldots + (z_r + p^\beta y_r)^s \equiv N(\mathrm{mod}\, p^\beta)$$

is solvable for all $y_2,\ldots,y_r$. Then the congruence

$$x_1^s \equiv N - \{(z_2 + p^\beta y_2)^s + \ldots + (z_r + p^\beta y_r)^s\}(\mathrm{mod}\, p^\beta)$$

is solvable in x_1 for each $y_2,\ldots,y_r$; $1 \le y_2,\ldots,y_r \le p^{m-\beta}$. By Lemma 5, the congruence

$$x_1^s \equiv N - \{(z_2 + p^\beta y_2)^s + \ldots + (z_r + p^\beta y_r)^s\} (\mathrm{mod}\, p^m)$$

is solvable and $x_1 \equiv z_1 (\mathrm{mod}\, p^\beta)$, therefore it follows that

$$T_\beta(p^m) \geq p^{(m-\beta)(r-1)} \tag{2.20}$$

thus , from (2.19) , (2.20) the Lemma 6 is proved . ■

Lemma 7. *The expression $G(z_1, \ldots, z_r; N)$ given in (2.7) is a positive real number non depending of $z_1, \ldots, z_r$.*

Proof . We will simplify the complicated expresion (2.7) by defining

$$G(z_1, \ldots, z_r; N) = \sum_{q=1}^{\infty} A(q)$$

$$A(q) = \sum_{\substack{a \leq q \\ (a,q)=1}} e\left(-\frac{a}{q}N\right) \prod_{j=1}^{r} \frac{(Q,q)}{q} \sum_{\substack{x_j=1 \\ x_j \equiv z_j (\mathrm{mod}\,(Q,q))}}^{q} e(\frac{a}{q}x_j^s).$$

It's trivial that $A(q) = \overline{A(q)}$ since the inner sum is the Ramanujan sum $c_q(x_j - z_j)$ and therefore it is an arithmetical function integer-valued. Moreover $A(q)$ is a multiplicative function and

$$\sum_{q=1}^{\infty} |A(q)| < +\infty.$$

It follows that

$$\sum_{q=1}^{\infty} A(q) = \prod_{p} (\sum_{\alpha=0}^{\infty} A(p^\alpha)) = \prod_{p \nmid Q} (1 + \sum_{\alpha=1}^{\infty} A(p^\alpha)) \prod_{p|Q} (1 + \sum_{\alpha=1}^{\infty} A(p^\alpha))$$

i) If $p \nmid Q$, then $(p^\alpha, Q) = 1$ for all $\alpha \geq 1$, and

$$A(p^\alpha) = \sum_{\substack{a \leq p^\alpha \\ (a,q)=1}} e(-\frac{aN}{p^\alpha}) \left(\frac{1}{p^\alpha} \sum_{x=1}^{p^\alpha} e(\frac{ax^s}{p^\alpha})\right)^r$$

$$|A(p^\alpha)| \leq \sum_{\substack{a \leq p^\alpha \\ (a,q)=1}} |\frac{1}{p^\alpha} \sum_{x=1}^{p^\alpha} e(\frac{ax^s}{p^\alpha})|^r \leq e^{2rs} p^{\alpha(-r/s+1)}.$$

Therefore

$$|\sum_{\alpha=1}^{\infty} A(p^\alpha)| \leq e^{2rs} p^{-3} \sum_{\alpha=1}^{\infty} p^{-\alpha(r/s-1)} \leq K e^{2rs} p^{-3} = c(r,s) p^{-3}$$

where $K = \sum_{\alpha=1}^{\infty} 2^{-\alpha(r/s-1)+3}$. If $p > c(r,s)$ then $1 + \sum_{\alpha=1}^{\infty} A(p^\alpha) > 1 - 1/p^2$, and for $p \leq c(r,s)$ we have $1 + \sum_{\alpha=1}^{\infty} A(p^\alpha) \geq p^{-(r-1)}$,([6] A.A.Karatzuba) . Thus we obtain

$$\prod_{p|Q}(1+\sum_{\alpha=1}^{\infty}A(p^{\alpha}))>\prod_{\substack{p\nmid Q\\ p>c(r,s)}}(1-\frac{1}{p^2})\prod_{\substack{p\nmid Q\\ p\le c(r,s)}}p^{-(r-1)}>\frac{1}{\zeta(2)}C=C\frac{6}{\pi^2}>0$$

ii) If $p|Q$ then exist β such that $p^{\beta}\parallel Q$. If $\alpha\le\beta$ then $(p^{\alpha},Q)=p^{\alpha}$, and $A(p^{\alpha})=p^{\alpha}-p^{\alpha-1}$. If $\alpha>\beta$ then $(p^{\alpha},Q)=p^{\beta}$, we have

$$A(p^{\alpha})=\sum_{\substack{a\le p^{\alpha}\\ (a,p)=1}}e(-\frac{aN}{p^{\alpha}})\prod_{j=1}^{r}p^{\beta-\alpha}\sum_{\substack{x_j=1\\ x_j\equiv z_j(\mathrm{mod}\,p^{\beta})}}^{p^{\alpha}}e(\frac{ax^s}{p^{\alpha}})=$$

$$=p^{(\beta-\alpha)r}\sum_{\substack{x_1,\dots,x_r=1\\ x_j\equiv z_j(\mathrm{mod}\,p^{\beta})}}^{p^{\alpha}}\sum_{a=1}^{p^{\alpha}}e\left(a\frac{(x_1^s+\dots+x_r^s-N)}{p^{\alpha}}\right)-$$

$$-p^{\beta r}\sum_{a=1}^{p^{\alpha-1}}e\left(-\frac{aN}{p^{\alpha-1}}\right)\prod_{j=1}^{r}\frac{p}{p^{\alpha}}\sum_{\substack{x_j=1\\ x_j\equiv z_j(\mathrm{mod}\,p^{\beta})}}^{p^{\alpha-1}}e\left(\frac{ax_j}{p^{\alpha-1}}\right).$$

Hence

$$A(p^{\alpha})=p^{\beta r}[p^{-\alpha(r-1)}T_{\beta}(p^{\alpha})-p^{-(\alpha-1)(r-1)}T_{\beta}(p^{\alpha-1})]$$

where $T_{\beta}(p^m)$, is the number of solutions of the congruence

$$x_1^s+\dots+x_r^s\equiv N(\mathrm{mod}\,p^m);\quad x_j\equiv z_j(\mathrm{mod}\,p^{\beta});\quad j=1,\dots,r;\quad m>\beta;$$

$T_{\beta}(p^m)$ is given in Lemma 6. If $m=\beta$, $T_{\beta}(p^{\beta})=1$, therefore , we obtain

$$\sum_{\alpha=\beta+1}^{m}A(p^{\alpha})=p^{\beta r}(p^{-m(r-1)}T_{\beta}(p^m)-p^{-\beta(r-1)})=\frac{p^{\beta r}}{p^{m(r-1)}}T_{\beta}(p^m)-p^{\beta}$$

And then $\forall m>\beta$:

$$\sum_{\alpha=0}^{m}A(p^{\alpha})=p^{\beta r}p^{-m(r-1)}T_{\beta}(p^m)=p^{\beta r}p^{-m(r-1)}p^{(r-1)(m-\beta)}=p^{\beta}.$$

Then

$$\prod_{p^{\beta}\parallel Q}(1+\sum_{\alpha=1}^{\infty}A(p^{\alpha}))=Q>0.$$

Thus we obtain

$$G(z_1,\dots,z_r;N)=G(N)=\sum_{q=1}^{\infty}A(q)>\frac{1}{\zeta(2)}CQ>0.$$

Note that $G(z_1,\dots,z_r;N)=G(N)$ is independient of $z_1,\dots z_r$ and a positive real numbers series and absolutely convergent . The Lemma 7 is proved and consecuently .

the following theorem is proved . ■

Theorem . *We suppose that* $r \geq c_3 s^2 \ln s$, $s > 2$ $P = [N^{\frac{1}{s}}]$,*and* $Q \leq P^c$ *with* $c \leq \min\{\frac{1}{4r}, \frac{1}{2^{s+4}}\}$. *Then we have the following asymptotic formula uniform on* Q.

$$R(N) = \frac{(\Gamma(1+\frac{1}{s}))^r}{\Gamma(r/s)} \frac{P^{r-s}}{Q^r} G(N) + O(\frac{P^{r-s-\frac{1}{2^{s+4}}}}{Q^r}) \quad \text{as} \quad N \to \infty$$

where $G(N) \geq c > 0$ *is a certain aritmethical function of* N .

REFERENCES

1] Arhipov,G.I. ,Karatzuba,A.A., Chubarikov,V.N.,*Multiple trigonometric sums.* Trudy Mat. Steklov **151** (1980); English translation Proc. Steklov Inst. Math **2 (151)** (1982)

2] Arhipov,G.I. ,Karatzuba,A.A., Chubarikov,V.N.,Special cases of the theory of multiple trigonometric sums , Math . USSR Izvestija , 23 (1984) 17-82

3] Ayoub,R, *An introduction to the analytic theory of numbers.* A.M.S. Providence , Rhode Island. (1963)

4] Hua Loo-Keng .*On a generalized Waring problem* .Proc. Lond. Math. Soc. **2 (43)** (1937) , 162-182 .

5] Hua Loo-Keng .*Additive theory of prime numbers* Amer. Math. Society (1966)

6] Karatzuba,A.A., *Fundamentos de la teoría analítica de los números.* Moscú: Edit. Mir, 1979

7] Minggao,Q.,Ding Ping., *On estimate of complete trigonometric sums.*Chin. Ann. of Math. **6B(1) ,** (1985) , 109-120 .

8] Mit'kin , D.A. *An estimate of the number of terms in Waring's problem for polynomials of general form* . Math. USSR Izv. Vol 29 (1987) , No. 2 371-406 .

9] Vinogradov,I.M., *The method of trigonometrical sums in the theory of numbers.*T.I.Steklov,**23,** translated from the Russian , revised and annoted by Davenport , A ; Roth , K.F. New York: Interscience.

Rev. Academia de Ciencias. Zaragoza. 48 (1993)

ON MIXED TRILATERAL GENERATING FUNCTIONS OF EXTENDED JACOBI POLYNOMIALS

Manik Chandra Mukherjee

Netajinagar Vidyamandir, Calcuta-700092. INDIA.

Abstract

In this paper the author establishes a teorem on mixed trilateral generating functions of extended Jacobi polynomials. Some special cases of the theorem have also discussed.

1. Introduction.

Putting $\theta = \frac{\lambda}{b-a}$ and $D = \frac{d}{dx}$, the extended Jacobi polinomial as defined in [1], is

$$F_n(\alpha,\beta;x) = \frac{(-1)^n}{n!}\,\theta^n\,(x-a)^{-\alpha}(b-x)^{-\beta}\,D^n[(x-a)^{n+\alpha}\,(b-x)^{n+\beta}] \qquad (1.1)$$

The object of this paper is to prove the following theorem in connection with the mixed trilateral generating relation which inturn yields the corresponding results involving Hermite, Laguerre, Bessel and Jacobi polynomials.

Theorem

If there exists a generating relation of the type

$$G(x,u,w) = \sum_{n=0}^{\infty} a_n\,F_n(\alpha,\beta;x)\;g_n(u)\;w^n \qquad (1.2)$$

and we put $\rho = \theta(x-a)$, then

$$(1-\lambda t)^{\alpha}[1-\rho t]^{-1-\alpha-\beta}\, G\left(\frac{x-\rho t}{1-\rho t},\ u,\ \frac{-zt}{1-\rho t}\right) = \sum_{n=0}^{\infty} t^n\, \sigma_n(z,u,x)$$

(1.3)

where

$$\sigma_n(z,u,x) = \sum_{k=0}^{n} (-1)^k\, a_k \binom{n}{k} F_n(\alpha-n+k,\beta;x)\, g_k(u)\, z^k)$$

The importance of our theorem lies the fact that one can get a large number of mixed trailateral generating relation from (1.3) by attributing different values to a_n in (1.2).

Proof of the Theorem:

Let

$$G(x,u,w) = \sum_{n=0}^{\infty} a_n\, F_n(\alpha,\beta;x)\, g_n(u)\, w^n \qquad (1.4)$$

Multiplying both sides of (1.4) by z^{α} and writing wy for w, we have

$$z^{\alpha}\, G(x,u,wy) = \sum_{n=0}^{\infty} a_n\, w^n\, [F_n(\alpha,\beta;x)\, y^n\, z^{\alpha}]\, g_n(u) \qquad (1.5)$$

Now we consider the operator

$$R = \theta\left[(x-a)(x-b)yz^{-1}\frac{\partial}{\partial x} - (x-a)y^2z^{-1}\frac{\partial}{\partial y} + (b-x)y\frac{\partial}{\partial z} - (1+\beta)(x-a)yz^{-1}\right]$$

such that

$$R[F_n(\alpha,\beta;x)\, y^n)z^{\alpha}] = -(n+1)\, F_{n+1}(\alpha-1,\beta;x)\, y^{n+1}\, z^{\alpha-1} \qquad (1.6)$$

The extended form of the group generated by R is given by

$$e^{wR} f(x,y,z) = \left[1+\frac{wy\rho}{\lambda}\right]^{-\beta-1} f\left(\frac{xz+bwy\rho}{z+wy\rho}, \frac{zy}{z+wy\rho}, \frac{z(z+\lambda wy)}{z+wy\rho}\right) \qquad (1.7)$$

Operating both sides of (1.5) by e^{wR}, we get

$$e^{wR}\, z^{\alpha}\, G(x,u,wy) = e^{wR}\left\{\sum_{n=0}^{\infty} a_n\, w^n\, [F_n(\alpha,\beta;x)\; y^n\, z^{\alpha}]\; g_n(u)\right\} \qquad (1.8)$$

Now from the last member of (1.8) we get

$$e^{wR}\, z^{\alpha}\, G(x,u,wy) = \left(1+\frac{wy^2\rho}{z\lambda}\right)^{-1-\alpha-\beta}\; z^{\alpha}\left(1+\frac{\lambda wy}{z}\right)^{\alpha}\; G\left(\frac{xz+bw\rho(b-a)}{z+wy\rho}\,,\right.$$
$$\left., u\,,\ \frac{wyz}{z+wy\rho}\right) \qquad (1.9)$$

From the right member of (1.8) we get

$$e^{wR}\left\{\sum_{n=0}^{\infty} a_n\, w^n\, [F_n(\alpha,\beta;x)\; y^n z^{\alpha}]\; g_n(u)\right\} = \sum_{n=0}^{\infty} S$$

where

$$S = \sum_{k=0}^{n} (a_n w^{n+k}/k!)\,(-1)^k\,(n+1)_k\, F_{n+k}(\alpha-k,\beta;x)\; y^{n+k}\, z^{\alpha-k}\, g_n(u)$$

Therefore, we obtain finally

$$e^{wR}\left\{\sum_{n=0}^{\infty} a_n\, w^n\, [F_n(\alpha,\beta;x)\; y^n z^{\alpha}]\; g_n(u)\right\} = \qquad (1.10)$$

$$= \sum_{n=0}^{\infty} -(wy/z)^n \sum_{k=0}^{n} (-z)^{n-k}(a_{n-k}/k!)(n-k+1)_k\, F_n(\alpha-k,\beta,x)\; g_{n-k}(u)\; z^{\alpha}$$

Equating the results (1.9) and (1.10) and putting $(-wy/z) = t$ we have

$$(1-\lambda t)^{\alpha}\,[1-\rho t]^{-1-\alpha-\beta}\; G\left(\frac{x-b\rho t(x-a)}{1-\rho t}\,,\ u\,,\ \frac{-zt}{1-\rho t}\right) = \sum_{n=0}^{\infty} t^n\, \sigma_n(x,u,z)$$

where

$$\sigma_n(x,u,z) = \sum_{k=0}^{n} (-1)^k\, a_k \binom{n}{k} F_n(\alpha-n+k,\beta;x)\; g_k(u)\; z^k)$$

Some special cases:

We now proceed to discuss some special cases.

Special case 1:

Putting $\alpha = \beta$, $b = -a = \sqrt{\alpha}$, $\lambda \to 2/\sqrt{\alpha}$, $\alpha \to \infty$, we get the following result involving Hermite polinomial.

Result 1:

If there exists a generating relation of the type

$$G(x,u,w) = \sum_{n=0}^{\infty} a_n \, [H_n(x)/n!] \, g_n(u) \, w^n$$

then

$$\exp(2xt-t^2) \, G(x-t \, , \, zt) = \sum_{n=0}^{\infty} a_n \, [H_n(x)/n!] \, \sigma_n(z,u) \, t^n$$

where

$$\sigma_n(z,u) = \sum_{k=0}^{n} a_k \binom{n}{k} g_k(u) \, z^k)$$

Special case 2 :

Putting $a = 0$, $\beta = b$, $\lambda = 1$, $b \to \infty$ and $L\{\mu,n\}(x) = L_n^{(\mu)}(x)$, we get the following result involving Laguerre polinomial.

Result 2: If

$$G(z,u,w) = \sum_{k=0}^{n} a_n \, L\{\alpha.n\}(x) \, g_n(u) \, w^n$$

then

$$\exp(-xt)(1+t)^{\alpha} \, G(x+xt \, , \, u \, , \, tz) = \sum_{n=0}^{\infty} t^n \, \sigma_n(z,u,x)$$

where

$$\sigma_n(z,u,x) = \sum_{k=0}^{n} a_k \binom{n}{k} L\{\alpha-n+k,n\}(x) \, g_k(u) \, z^k)$$

Special case 3 :

Putting $b = -a = \lambda = 1$, $\alpha = \nu-\varepsilon-1$, $\beta = \varepsilon - 1$, replacing t by $s\omega/\varepsilon$, x by $1 + (2x\varepsilon/s)$, and $\varepsilon \to \infty$, we get the following result involving Bessel polinomial.

Result 3 : If

$$G(x,u,w) = \sum_{n=0}^{\infty} a_n \, Y_n(x,\alpha,s) \, g_n(u) \, w^n$$

then

$$\exp(t)(1 - xt/s)^{1-\nu} \, G\left(\frac{x}{1-xt/s} \, , \, u \, , \, \frac{zt}{1-xt/s}\right) = = \sum_{n=0}^{\infty} (t^n/n!) \, \sigma_n(z,u,x)$$

where

$$\sigma_n(z,u,x) = \sum_{k=0}^{n} a_k \binom{n}{k} Y\{x,\alpha-n+k,s\} \, g_k(u) \, z^k)$$

Special case 4 :

Putting $-a = b = 1$, $\lambda = 1$, $q = \frac{t}{2}(x+1)$, $P\{(\beta,\alpha),n\} = P_n^{(\beta,\alpha)}$, we get the following result involving Jacobi polynomial.

Result 4 : If

$$G(x,u,w) = \sum_{n=0}^{\infty} a_n \, P\{(\beta,\alpha),n\}(x) \, g_n(u) \, w^n$$

then

$$(1-t)^{\alpha}\{1 - q\}^{-1-\alpha-\beta} \, G\left(\frac{x-q}{1-q} \, , \, u \, , \, \frac{-zt}{1-q}\right) = \sum_{n=0}^{\infty} t^n \, \sigma_n(z,u,x)$$

where

$$\sigma_n(z,u,x) = \sum_{k=0}^{n} (-1)^k \, a_k \binom{n}{k} P\{(\beta,\alpha-n+k),n\}(x) \, g_k(u) \, z^k)$$

References

1. K.R.Patil and N.K. Thakare: *Operational formulas and generating functions in the unified form for the classical orthogonal polinomials*. The Math. Student, **45**(1), 1977, pag. 41-51.

2. L. Weisner: *Group-theoretic Origin of certain Generating Functions*. Pacific J. Math. **5**, 1955, pag. 1033-1039.

Rev. Academia de Ciencias. Zaragoza. 48 (1993)

ON S-ORDER OF CONVERGENCE

Don Chen
Department of Mathematical Sciences
University of Arkansas, Fayetteville, USA.
Ioannis Argyros
Department of Mathematics, Cameron University
Lawton, Oklahoma 73505, USA.

Abstract.
We discuss the S-order convergence which was introduced by the first author in [2,3]. Σe classify the previous R-order as well the Q-order as a N-order (necessary order. We obtain some new properties of a fourth-(S) order Halley-Werner method which is third-(N) order [2]. We list the efficiency index (E.I.) table under senses of S-order and R-order (or Q-order) for well-known methods. Also we will give a numerical example to show that the definition of S-order is consistent with the numerical testing.

1.Introduction.

The definition of order of convergence in nonlinear approximation theory is less considered than of convergence theory [1,2,3,4,5,6,7,8]. The most people way for defining the order of convergence is by a natural estimate of the form:

$$\|x_{n+1} - x^*\| \leq c\|x_n - x^*\|^p, \quad n \geq 0.$$

Throughout this paper we will assume that the sequences $\{X_n\},\{Y_n\}$ introduced here belong in a metric space for all $n \geq 0$.

Ortega and Rheinboldt classified this idea and defined two strict definitions of order which are called R-order and Q-order [5]. It is well known that there are two different versions for the convergence of Newton's method. One is a necessary conditions and the other is the famous Ostrowski-Kantorovich theorem (sufficient conditions for convergence) [4,7]. But the R-order or Q-order constitute a necessary condition for finding the order which we classify it as N-order. It seems to us that the condition of order of convergence is not consistent with the condition of convergence. In this paper, our theorem essentially answers (partially) to the following question: Is it possible to find sufficient order (S-order) from the characteristics of iterations

2. The definition of S-order.

First let us recall the definition of S-order from [2,3].

Definition 2.1. (S-order) Let g(t) be a testing function of order 2. Assume that $g(t) = (K/2)t^2 - (1/b)t + (h/b)$, K, b, h , and $h = Kbh < (1/2)$ are positive real numbers. A sequence of iterations (one step or multistep iterations without memory) defined in a metric space is said to convergence with the order $p \geq 1$ to a point X^* if in

One step case:

$$E(g(t_{n+1}),\ t_n, t_{n+1}) = g(t_{n+1}) - C(t_n,\ t_{n+1})(t_{n+1} - t_n)^p = 0$$

Multipoint step case:

$$E(g(t_{n+1}),\ t_n, s_n, t_{n+1}) = g(t_{n+1}) - C(t_n,\ s_n)(s_n - t_n)^p = 0$$

for some $c > 0$ and for all $n \geq 0$, where

$$E(P(X_{n+1}),\ X_n,\ Y_n, X_{n+1}) = P(X_{n+1}) - R(X_n,\ Y_n,\ X_{n+1})$$

with being R the Ostrowski-Kantotovich representation of $P(X_{n+1})$.

Definition 2.2. The asymptotic error constant $C(t^*)$ is defined as follows:

One step case:
$$C(t^*) = \lim_{n \to \infty} \frac{g(t_{n+1})}{(t_{n+1} - t_n)^p};$$

Multistep case:
$$C(t^*) = \lim_{n \to \infty} \frac{g(t_{n+1})}{(s_n - t_n)^p},$$

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

Now we are ready to apply the definition of S-order to some well-known methods.

a. Newton's method.

It is well known that Newton's method is given by

$$X_{n+1} = X_n - P'(X_n)^{-1}P(X_n),\ n \geq 0. \qquad (2.3)$$

First we have to find the Ostrowski-Kantorovich representation of $P(X_{n+1})$. From the approximation for all $n \geq 0$,

$$P(X_{n+1}) = P(X_{n+1}) - P(X_n) - P'(X_n)(X_{n+1} - X_n) + P(X_n) + P'(X_n)(X_{n+1} - X_n),$$

since

$$P(X_n) + P'(X_n)(X_{n+1} - X_n) = 0,$$

and therefore we have

$$\begin{aligned} P(X_{n+1}) &= P(X_{n+1}) - P(X_n) - P'(X_n)(X_{n+1} - X_n) \\ &= \int_0^1 P''(X_n + t(X_{n+1} - X_n))(1 - t)dt(X_{n+1} - X_n)^2 \end{aligned} \tag{2.4}$$

Replacing P(X) by g(t) and have

$$\begin{aligned} g(t_{n+1}) &= \int_0^1 g''(t_n + t(t_{n+1} - t_n))(1 - t)dt(t_{n+1} - t_n)^2 \\ &= \int_0^1 K(1 - t)dt(t_{n+1} - t_n)^2 = \frac{K}{2}(t_{n+1} - t_n)^2 . \end{aligned} \tag{2.5}$$

So, we get

$$R(t_n, t_{n+1}) = \frac{K}{2}(t_{n+1} - t_n)^2$$

$$E(g(t_{n+1}), t_n, t_{n+1}) = g(t_{n+1}) - \frac{K}{2}(t_{n+1} - t_n)^2$$

Thus, we finally have $p_S = 2 = p_N$, and

$$C_N(t^*) = \lim_{n \to \infty} \frac{g(t_{n+1})}{(t_{n+1} - t_n)^2} = \frac{K}{2} . \tag{2.6}$$

b. Chebyshev method.

The Chebyshev method is described by

$$\begin{aligned} Y &= X_n - \frac{P(X_n)}{P'(X_n)} \\ X_{n+1} &= Y_n - \frac{1}{2}\frac{P''(X_n)}{P'(X_n)}(Y_n - X_n)^2 , \quad n \geq 0. \end{aligned} \tag{2.7}$$

It is easy to see that we have the following Ostrowski- Kantotovich representation for all $n \geq 0$:

$$\begin{aligned} P(X_{n+1}) &= \int_0^1 P''(Y_n + t(X_{n+1} - Y_n))(1 - t)dt(X_{n+1} - Y_n)^2 \\ &\quad - \frac{P''(X_n)}{P'(X_n)}(Y_n - X_n)^3 \int_0^1 P''(X_n + t(Y_n - X_n))dt \\ &\quad + \frac{1}{2}(Y_n - X_n)^2 \int_0^1 [2P''(X_n + t(Y_n - X_n))(1 - t) - P''(X_n)]dt. \end{aligned} \tag{2.8}$$

Replacing P(X) by g(t), we obtain

$$g(t_{n+1}) = \int_0^1 g''(s_n + t(t_{n+1} - s_n))(1 - t)dt(t_{n+1} - s_n)^2$$

$$- \frac{g''(t_n)}{2g'(t_n)} (s_n - t_n)^2 \int_0^1 g''(t_n + t(s_n - t_n))dt$$

$$+ \frac{1}{2} (s_n - t_n)^2 \int_0^1 [2g''(t_n + t(s_n - t_n))(1 - t) - g''(t_n)]dt$$

$$= \left[\frac{K^3}{8g'(t_n)^2} (s_n - t_n) - \frac{K^2}{2g'(t_n)}\right](s_n - t_n)^3$$

$$= c_C(t_n, s_n)(s_n - t_n)^3 \text{ , for all } n \geq 0.$$

Thus, we get $p_S = 3 = p_N$, where

$$c_C(t^*) = \lim_{n \to \infty} \frac{g(t_{n+1})}{(s_n - t_n)^3} = \frac{K^2\beta}{2\sqrt{1 - 2h}} . \tag{2.10}$$

c. Halley method.

The halley method is defined by

$$Y_n = X_n - \frac{P(X_n)}{P'(X_n)}$$

$$X_{n+1} = Y_n - \frac{1}{2} \frac{\frac{P''(X_n)}{P'(X_n)} (Y_n - X_n)^2}{1 + \frac{1}{2} \frac{P''(X_n)}{P'(X_n)} (Y_n - X_n)}, \text{ for } n \geq 0. \tag{2.11}$$

We find the Ostrowski-Kantorovich representation of $P(X_{n+1})$ for all $n \geq 0$ as follows:

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

$$- \frac{\frac{1}{2} \frac{P''(X_n)}{P'(X_n)} (Y_n - X_n)^3 \int_0^1 P''(Y_n + t(X_{n+1} - Y_n))tdt}{1 + \frac{1}{2} \frac{P''(X_n)}{P'(X_n)} (Y_n - X_n)} \tag{2.12}$$

$$+ \frac{\frac{1}{2} (Y_n - X_n)^2 \int_0^1 [2P''(X_n + t(Y_n - X_n))(1 - t) - P''(X_n)]dt}{1 + \frac{1}{2} \frac{P''(X_n)}{P'(X_n)} (Y_n - X_n)} .$$

So, putting $H = \frac{K}{2g'(t_n)}$, and $G = H(s_n - t_n)$, we obtain

$$g(t_{n+1}) = \left[\frac{2G}{2 + G} - \frac{1}{1 + G}\right]\frac{GK(s_n - t_n)^2}{2} = c_H(t_n, s_n)(s_n - t_n) \qquad (2.13)$$

Hence, we deduce that $p_S = 3 = p_N$, and $c_H(t^*) = \frac{K^2\beta}{4\sqrt{1-2h}}$

From Traub's theory point of view [9,10], the maximum order of convergence of Chebyshev as well as Halley methods by using triplet information (f, f', f'') at the same point is 3. Now we will show that it is no true under the sense of S-order.

d. A Halley-Werner Type Fourth-order method.

We define the iteration

$$\begin{aligned} Y_n &= X_n - \frac{P(X_n)}{P'(X_n)} \\ X_{n+1} &= Y_n - \frac{\frac{1}{2}\frac{P''(X_n)}{P'(X_n)}(Y_n - X_n)^2}{1 + \frac{P''(X_n)}{P'(X_n)}(Y_n - X_n)}, \quad \text{for } n \geq 0. \end{aligned} \qquad (2.14)$$

Then, the Ostrowski-Kantorovich representation for all $n \geq 0$, is

$$\begin{aligned} P(X_{n+1}) &= \int_0^1 P''(Y_n + t(X_{n+1} - Y_n))(1 - t)dt(X_{n+1} - Y_n)^2 \\ &- \frac{\frac{P''(X_n)}{2P'(X_n)}(Y_n - X_n)^3\int_0^1 P''(X_n + t(Y_n - X_n))(2t - 1)dt}{1 + \frac{P''(X_n)}{P'(X_n)}(Y_n - X_n)} \\ &+ \frac{\frac{1}{2}(Y_n - X_n)^2\int_0^1\left[2P''(X_n + t(Y_n - X_n))(1 - t) - P''(X_n)\right]dt}{1 + \frac{P''(X_n)}{P'(X_n)}(Y_n - X_n)} \end{aligned} \qquad (2.15)$$

Notice that

$$\int_0^1 g''(t_n + t(s_n - t_n))(2t - 1)dt = \int_0^1 K(2t - 1)dt = 0$$

and

$$\int_0^1 \left[2g''(t_n + t(s_n - t_n))(1 - t) - g''(t_n)\right]dt = 0$$

Therefore, we deduce

$$g(t_{n+1}) = \int_0^1 g''(s_n + t(t_{n+1} - s_n))(1 - t)dt\,(t_{n+1} - s_n)^2$$

$$= \frac{K}{2}(t_{n+1} - s_n)^2 = \frac{K^3/8g'(t_n)^2}{\left[1 + \frac{K}{g'(t_n)}(s_n - t_n)\right]^2}(s_n - t_n)^4 \qquad (2.16)$$

$$= C_{H\text{-}W}(t_n, s_n)(s_n - t_n)^4 .$$

Hence, we get $p_S = 4 > 3 = p_N$, and $c_{H\text{-}W}(t^*) = \frac{K^3\beta^2}{8(1-2h)}$

3. Some Properties of the Asymptotic Error Constants and E.I.

Table

In this section, we introduce a new definition to classify general iterative methods which is called the equivalent class.

Definition 3.1.

Two iterative functions ψ_1 and ψ_2 are said to be in the equivalent class if the ratio of their asymptotic error constants is a non-zero number γ. That is,

$$\psi_1 \equiv \psi_2 \quad \text{iff} \quad \frac{C_{\psi_1}(t^*)}{C_{\psi_2}(t^*)} = \gamma. \qquad (3.2)$$

Definition 3.3.

Two iterative functions ψ_1 and ψ_2 are said to be not in the equivalent class with order n if the ratio of their asymptotic error constants is $K^n\beta^n\gamma/(1-2h)^{n/2}$, where γ is a independent non-zero

Theorem 3.4: According to the above definitions we have the following conclusions: (i) Chebyshev and Halley methods are in the equivalent class; (ii) Newton and Chebyshev (or Halley) methods are not in the equivalent class with order one; (iii) Newton and Halley-Werner type methods are not in the equivalent class with order two; (iv) Chebyshev (or Halley) and Halley-Werner methods are not in the equivalent class with order one.

Proof: Let us recall that

$$C_N = \frac{K}{2},\ C_C = \frac{K^2\beta}{2\sqrt{1 - 2h}},\ C_H = \frac{K^2\beta}{4\sqrt{1 - 2h}},\ C_{H-W} = \frac{K^3\beta^2}{8(1 - 2h)}$$

Therefore, we obtain

$$\frac{C_C}{C_H} = 2,\ \frac{C_C}{C_N} = \frac{K\beta}{\sqrt{1 - 2h}},\ \frac{C_H}{C_N} = \frac{K\beta}{2\sqrt{1 - 2h}},$$

$$\frac{C_{H-W}}{C_n} = \frac{K^2\beta^2}{4(1 - 2h)},\ \frac{C_{H-W}}{C_C} = \frac{K\beta}{4\sqrt{1 - 2h}}.$$

that completes the proof of the theorem. We now list two different E.I. tables.

Table 1

	Newton(2.3)	Chebyshev(2.7)	Halley(2.11)	Halley-Werner(2.14)
EI(R)	1.414	1.442	1.442	1.442
EI(S)	1.414	1.442	1.442	1.587

4. The Numerical Example: In this section, we present two numerical examples to illustrate the previous theoretical conclusions. We consider the function $f(X) = X^3 - 2X - 5$. Denote $E_1 = |X_1 - X^*|$. Then, we have the following results:

Table 2

	Newton(2.3)	Chebyshev(2.7)	Halley(2.11)	Halley-Werner(2.14)
X_0	2	2	2	2
X_1	2.1	2.094	2.0943396	2.0946429
E_1	$0.5449*10^{-2}$	$0.551*10^{-3}$	$0.211*10^{-3}$	$0.91*10^{-4}$

REFERENCES

[1] W. Burmeister and J.W. Schmidt: On the R-order of Coupled Sequences, Part II and III, Computing, 29 (1982), 73-81; 30 (1984), 157-169.

[2] Dong Chen: On a New Definition of order of Convergence in General Iterative Methods I: One Point Iterations, Submitted.

[3] Dong Chen: On a New Definition of order of Convergence in General Iterative Methods II: Multipoint Iterations, Submitted.

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

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

[6] F.A. Potra: On O-order and R-order of Convergence, J. Optimization Theory and Applications, 63: 3 (1989), 415-431.

[7] A.M. Ostrowaki: Solution of Equations and Systems of Equations, Third Edition, Academic Press, New York, 1973.

[8] J.W. Schmidt: On the R-order of Coupled Sewquences, Computing, 26 (1981), 333-342.

[9] J.F. Traub: Iterative Methods for Solution of Equations, Englewood-Cliffs, Prentice Hall, New Jersey, 1964.

[10] J.F. Traub and H. Wozniakowski: A General Theory of Optimal Algorithms, Academic Press, New York, 1980.

Rev. Academia de Ciencias. Zaragoza. 48 (1993)

ON THE A POSTERIORI ERROR BOUNDS FOR A CERTAIN ITERATION UNDER ZABREJKO-NGUEN ASSUMPTIONS

by

IOANNIS K. ARGYROS
Cameron University
Department of Mathematics
Lawton, OK 73505-6377, U. S. A.

Abstract. Recently Yamamoto [8] gave a posteriori error bounds for the Zinceko iteration for solving nonlinear equations. Using Zabrejko and Nguen assumptions [9], he showed that his estimates are sharper than Miel-type bounds [3]. Under similar assumptions we show that our a posteriori error bounds are sharper than Yamamoto's.

Key words and phrases. Banach space, Newton's method, nonlinear operator equation.

(1980) A.M.S. classification codes: 49D15, 47H17, 65.

I. INTRODUCTION. Let X and Y be Banach spaces and $U(x_0, R)$ be the closed ball with center x_0 and radius R in X. Suppose that the operators F and G are defined on $U(x_0, R)$, with values in Y such that F is Fréchet differentiable at every interior point of $U(x_0, R)$, $F'(x_0)^{-1}$ exists and

$$\|F'(x_0)^{-1}(F'(x+h)-F'(x))\| \leq A(r,\|h\|),\ x \in U(x_0,r),\ 0 \leq r \leq R,\ 0 \leq \|h\| \leq R-r \tag{1}$$

while G satisfies the condition

$$\|F'(x_0)^{-1}(G(x+h)-G(x))\| \leq B(r,\|h\|),\ x \in U(x_0,r),\ 0 \leq r \leq R,\ 0 \leq \|h\| \leq R-r\ . \tag{2}$$

Here A, B are positive, continuous functions of two variables such that if one of the va iables is fixed then A, B are non-decreasing functions of the other on the interval [0, R]. Moreover, the following are true:

(a) the function $\frac{\partial A(0,t)}{\partial t}$ is postiive, continuous and non-decreasing on [0, R - r] with A(0, 0) = 0;

(b) B is linear in the second variable and the function $\frac{\partial B(R,t)}{\partial t}$ is positive, continuous and non-decreasing on [0, R - r].

We are concerned with approximating a solution x^* of the equation

$$F(x) + G(x) = 0 \text{ in } U(x_0, R), \tag{3}$$

using Newton's approximations given by

$$x_{n+1} = x_n - F'(x_n)^{-1}(F(x_n) + G(x_n)), \quad n = 0, 1, 2, \cdots. \tag{4}$$

Zabrejko and Nguen in [9] gave sufficient conditions for the existence and uniqueness of solution x^* in $U(x_0, R)$, as well as error estimates on the distances $\|x^* - x_n\|$ and $\|x_{n+1} - x_n\|$ when

$$A(r, t) = k(r)t \tag{5}$$

and

$$B(r, t) = \epsilon(r)t \tag{6}$$

where k(r) and $\epsilon(r)$ are non-decreasing functions on the interval [0, R]. Yamamoto in [8] gave a posteriori error bounds for the iteration (4) following a suggestion given in [9].

Using the forms (1) and (2) instead of (5) and (6) we showed in [1] that our error estimates on the distances $\|x_n - x^*\|$ and $\|x_n - x_{n+1}\|$ are sharper than the ones given by Zabrejko and Nguen in [9].

In this paper we provide some a posteriori error bounds for iteration (4) which are sharper than the ones given by Yamamoto in [8].

A simple example is also provided where our results compare favorably with the ones obtained by Yamamoto in [8].

II. PRELIMINARIES. To justify the claim made in the introduction we must reproduce some of the results and notation given in [1], [8] and [9] respectively.

Set

$$a = \|F'(x_0)^{-1}(F(x_0) + G(x_0))\| ,$$

$$\omega(r) = \int_0^r k(t)dt ,$$

$$\varphi(r) = a - r + \int_0^r \omega(t)dt ,$$

$$\varphi_\gamma(r) = (a - r)(1 - A(0, r)) + \int_0^r A(r, t)dt ,$$

$$\psi(r) = \int_0^r \epsilon(t)dt ,$$

$$\psi_\gamma(t) = B(R, t) ,$$

$$\chi(r) = \varphi(r) + \psi(r) ,$$

$$\chi_\gamma(r) = \varphi_\gamma(r) + \psi_\gamma(r), \quad \text{for all} \quad r \in [0, R] .$$

Moreover, set

$$r_n = \|x_n - x_0\|, \; k_n(r) = k(r_n + r), \; \epsilon_n(r) = \epsilon(r_n + r) ,$$

$$a_n = \|x_{n+1} - x_n\|, \; b_n = (1 - \omega(r_n))^{-1},$$

$$U_n(r) = u_n(r) - r + a_n , \tag{7}$$

$$V_n(r) = v_n(r) - r + a_n , \tag{8}$$

and

$$W_n(r) = v_n(r) - r + \Delta\rho_n , \tag{9}$$

where

$$u_n(r) = b_n\int_0^r [\int_0^t k(r_n + s)ds + \epsilon(r_n + t)]dt ,$$

$$v_n(r) = c_n\int_0^r [\int_0^t k(\rho_n + s)ds + \epsilon(\rho_n + t)]dt ,$$

$$c_n = (1 - \omega(\rho_n))^{-1}, \; \Delta\rho_n = \rho_{n+1} - \rho_n$$

and

$$\rho_{n+1} = \rho_n - \frac{\chi(\rho_n)}{\varphi'(\rho_n)}, \quad \rho_0 = 0, \quad n = 0, 1, 2, \cdots . \tag{10}$$

Yamamoto showed in [8, p. 989] that the equation

$$U_n(r) = 0, \quad n = 0, 1, 2, \cdots \tag{11}$$

has a unique solution ρ_n^* in $[0, \rho^* - \rho_n]$ where $\rho_0^* = \rho^* = \lim_{n\to\infty} \rho_n$, provided that $a_n > 0$, $n = 0, 1, 2, \cdots$.

We can state Proposition 4 in [9] and Theorem 2 in [8] in a combined form as follows:

THEOREM 1. Suppose that the function $\chi(r)$ has a unique zero ρ^* in the interval $[0, R]$ and $\chi(R) \leq 0$. Then equation (3) has a solution x^* in the ball $U(x_0, \rho^*)$, which is unique in $U(x_0, R)$. The iterates (4) are defined for all n, belong to $U(x_0, \rho^*)$ and satisfy the estimates

$$\|x_{n+1} - x_n\| \leq \rho_{n+1} - \rho_n$$

and

$$\begin{aligned} \|x^* - x_n\| &\leq \rho_n^*, && n = 0, 1, 2, \cdots \\ &\leq (\rho^* - \rho_n)a_n/\Delta\rho_n, && n = 0, 1, 2, \cdots \\ &\leq (\rho^* - \rho_n)a_{n-1}/\Delta\rho_{n-1}, && n = 1, 2, \cdots \\ &\leq \rho^* - \rho_n, && n = 0, 1, 2, \cdots, \end{aligned}$$

where the sequence $\{\rho_n\}$ is given by (10).

Furthermore, we must define

$$A_n(r, t) = A(r, r_n + t), \quad B_n(R, t) = B(R, t + r_n),$$

$$\bar{b}_n = (1 - A(0, r_n))^{-1}, \quad \bar{c}_n = (1 - A(0, \rho_n))^{-1},$$

$$\bar{U}_n(r) = \bar{u}_n(r) - r + a_n, \tag{12}$$

$$\bar{V}_n(r) = \bar{v}_n(r) - r + a_n, \tag{13}$$

and

$$\bar{W}_n(r) = v_n(r) - r + \Delta\bar{\rho}_n, \tag{14}$$

where

$$\bar{u}_n(r) = \bar{b}_n[\int_0^r A(r, t + r_n)dt + B(R, r_n + r)] ,$$

$$\bar{v}_n(r) = \bar{c}_n[\int_0^r A(r, t + \bar{\rho}_n)dt + B(R, \bar{\rho}_n + r)] ,$$

$$\Delta\bar{\rho}_n = \bar{\rho}_{n+1} - \bar{\rho}_n$$

and

$$\bar{\rho}_{n+1} = \bar{\rho}_n - \frac{\chi_\gamma(\bar{\rho}_n)}{\varphi'_\gamma(\bar{\rho}_n)} , \quad \bar{\rho}_0 = 0 , \quad n = 0, 1, 2, \cdots . \tag{15}$$

We will show that the equation

$$\bar{U}_n(r) = 0, \quad n = 1, 2, \cdots \tag{16}$$

has a unique solution $\bar{\rho}_n^*$ in $[0, \bar{\rho}^* - \bar{\rho}_n]$ where $\bar{\rho}_0^* = \bar{\rho}^* = \lim_{n\to\infty} \bar{\rho}_n$ was proven in Theorem 2 and Proposition 1 of [1]. In particular, we showed:

THEOREM 2. Suppose that the equation $\chi_\gamma(r)$ has a unique zero $\bar{\rho}^*$ in the interval $[0, R]$, $1 - A(0, R) > 0$ and $\chi_\gamma(R) \le 0$.

Then

(a) the sequence $\{\bar{\rho}_n\}$, $n = 0, 1, 2, \cdots$ given by (15) is monotonically increasing and converges to $\bar{\rho}^*$.

(b) The iterates generated by (4) are well-defined in $U(x_0, \bar{\rho}^*)$ for all n, and satisfy the estimates

$$\|x_{n+1} - x_n\| \le \bar{\rho}_{n+1} - \bar{\rho}_n \le \rho_{n+1} - \rho_n , \quad n = 0, 1, 2, \cdots \tag{17}$$

and

$$\|x_n - x^*\| \le \bar{\rho}^* - \bar{\rho}_n \le \rho^* - \rho_n , \quad n = 0, 1, 2, \cdots \tag{18}$$

provided that

$$A(r, t) \le \omega(t), \ 0 \le r \le R, \ 0 \le t \le R - r , \tag{19}$$

$$\psi_\gamma(r) \le \psi(r) , \quad 0 \le r \le R , \tag{20}$$

and

$$\rho^* - a \le (\bar{\rho}_2 - a)(1 - A(0, R)) . \tag{21}$$

We can now justify the claim made at the introduction.

THEOREM 3. Under the assumptions of Theorems 2 and 3 we have

(a) the following estimates are true

$$\|x^* - x_n\| \le \bar{\rho}_n^* , \qquad n = 0, 1, 2, \cdots$$

$$\le (\bar{\rho}^* - \bar{\rho}_n)a_n/\Delta\bar{\rho}_n , \qquad n = 0, 1, 2, \cdots$$

$$\le (\bar{\rho}^* - \bar{\rho}_n)a_{n-1}/\Delta\bar{\rho}_{n-1} , \qquad n = 1, 2, \cdots$$

$$\le \bar{\rho}^* - \bar{\rho}_n , \qquad n = 0, 1, 2, \cdots .$$

(b) Moreover, the following are true

$$\bar{\rho}_n^* \le \rho_n^* , \qquad n = 0, 1, 2, \cdots$$

$$(\bar{\rho}^* - \bar{\rho}_n)/\Delta\bar{\rho}_n \le (\rho^* - \rho_n)/\Delta\rho_n , \qquad n = 0, 1, 2, \cdots$$

$$(\bar{\rho}^* - \bar{\rho}_n)/\Delta\bar{\rho}_{n-1} \le (\rho^* - \rho_n)/\Delta\rho_{n-1} , \qquad n = 1, 2, \cdots$$

and

$$\bar{\rho}^* - \bar{\rho}_n \le \rho^* - \rho_n , \qquad n = 0, 1, 2, \cdots .$$

PROOF. The proof of the theorem will be based on the principles developed by Zabrejko-Nguen in [9], Yamamoto in [7], [8] and Argyros in [1].

Using the identity

$$x^* - x_{n+1} = -F'(x_n)^{-1}[\int_0^1 F'(x_n + \theta(x^* - x_n)) - F'(x_n)](x^* - x_n)d\theta$$

$$+ (G(x^*) - G(x_n))] ,$$

and the estimates (1), (2) we obtain with $e_n = \|x^* - x_n\|$

$$e_{n+1} \le \|F'(x_n)^{-1}F'(x_0)\|[\int_0^1 \|F'(x_0)^{-1}(F'(x_n + \theta(x^* - x_n)) - F'(x_n))\|$$

$$\cdot\|x^* - x_n\|d\theta + \|F'(x_0)^{-1}(G(x^*) - G(x_n))\|]$$

$$\le \bar{b}_n[\int_0^1 A_n(0, \theta e_n)e_n d\theta + B_n(0, e_n)]$$

$$\le \bar{b}_n[\int_0^{e_n} A(e_n, t + r_n)dt + B(R, e_n + r_n)] = \bar{u}_n(e_n)$$

$$\leq \bar{c}_n[\int_0^{e_n} A(e_n,\ t + \rho_n)dt + B(R,\ e_n + \rho_n)] = \bar{v}_n(e_n) .$$

By the definition of $\bar{U}_n$, $\bar{V}_n$ and $\bar{W}_n$ we have

$$\bar{U}_n(r) \leq \bar{V}_n(r) \leq \bar{W}_n(r)$$

for all $r \in [0,\ R - r_n]$.

To avoid repetitions we show as in [8, p. 991] that the equation $\bar{U}_n(r) = 0$ has a solution $\bar{\rho}_n^*$ in $[0,\ \bar{\rho}^* - \bar{\rho}_n]$ which is unique in $[0,\ R - r_n]$.

Moreover, let

$$q_n = (\bar{\rho}^* - \bar{\rho}_n)a_n/\Delta\bar{\rho}_n .$$

Then

$$\bar{V}_n(q_n) = \bar{c}_n^{-1}[\int_0^{q_n} A(q_n,\ t + \bar{\rho}_n)dt + B(R,\ q_n + \bar{\rho}_n)] - q_n + a_n$$

$$= \bar{c}_n^{-1}\frac{a_n}{\Delta\bar{\rho}_n}[\int_0^{\bar{\rho}^* - \bar{\rho}_n} A(q_n,\ t + \bar{\rho}_n)dt + B(R,\ \bar{\rho}^*)] - \frac{(\bar{\rho}^* - \bar{\rho}_n)}{\Delta\bar{\rho}_n} + a_n$$

$$= \frac{a_n}{\Delta\bar{\rho}_n}(\bar{w}_n(\bar{\rho}^* - \bar{\rho}_n) - \Delta\bar{\rho}_n) + a_n$$

$$= \frac{a_n}{\Delta\bar{\rho}_n}\bar{w}_n(\bar{\rho}^* - \bar{\rho}_n) = 0 .$$

That is

$$\bar{U}_n(q_n) \leq 0 .$$

Since,

$$e_n - a_n \leq e_{n+1} \leq u_n(e_n)$$

we obtain

$$\bar{U}_n(e_n) \geq 0 .$$

By

$$\bar{U}_n(d_n) \leq 0$$

we get

$$e_n \leq \bar{\rho}_n^* \leq q_n \leq \bar{\rho}^* - \bar{\rho}_n . \qquad (22)$$

Similarly we show that

$$a_n \leq (\Delta\bar{\rho}_n/\Delta\bar{\rho}_{n-1})a_{n-1} , \quad n = 1, 2, \cdots . \tag{23}$$

The estimates in part (a) of the theorem now follow from (22) and (23).

(b) Using hypotheses (19), (20) and (21) we obtain

$$\bar{U}_n(r) \leq U_n(r) \quad \text{for all} \quad r \in [0, R] . \tag{24}$$

In particular,

$$\bar{U}_n(\rho_n^*) \leq U_n(\rho_n^*) = 0 .$$

That is

$$\bar{\rho}_n^* \leq \rho_n^* , \quad \text{for all} \quad n = 0, 1, 2, \cdots .$$

The rest of the estimates in part (b) follow similarly using (17), (18) and (24).

That completes the proof of the theorem.

As in [8, p. 993] using the Gragg and Tapia technique [3] and the proof of Theorem 3 we can easily provide lower bounds on the distances $\|x_n - x^*\|$ which are sharper than the ones given by Yamamoto in [8]. We leave the details to the motivated reader.

III. APPLICATIONS. We now provide an example where our estimates compare favorably with the ones in [8]. Let us assume for simplicity that $G = 0$ and consider the real equation

$$F(x) = e^x - e \quad \text{on} \quad [x_0, R]$$

with $x_0 = .9$ and $R = .11$. We set

$$k(r) = e^r ,$$

and

$$A(r, t) = e^r t .$$

Then using the definitions it can easily be seen that

$$\rho_0^* = .10761178 \quad \text{and} \quad \bar{\rho}_0^* = .10752104.$$

It is now simple calculus to show that the hypotheses of Theorem 3 are satisfied. Therefore, the conclusions apply.

REFERENCES

[1] I.K. Argyros. On the solution of equations with nondifferentiable operators and Ptâk error estimates.

[2] M. Balazs and G. Goldner. On the method of the cord and on a modification of it for the solution of nonlinear operator equations. Stud. Cerc. Mat. 20, (1968), 981-990.

[3] W.B. Gragg and R.A. Tapia. Optimal error bounds for the Newton-Kantorovich Theorem. S.I.A.M. J. Numer. Anal. 11 (1974), 10-13.

[4] F.A. Potra and V. Ptâk. Sharp error bounds for Newton's process. Numer. Math. 34, (1980), 63-72.

[5] W.C. Rheinboldt. A unified convergence theory for a class of iterative processes. S.I.A.M. J. Numer. Anal. 5 (1968), 42-63.

[6] J.W. Schmidt. Unter Fehrerschranker für regular-falsi-verfahren. Period. Math. Hung. 9, (1978), 241-247.

[7] T. Yamamoto. A method for finding sharp error bounds for Newton's method under the Kantorovich assumptions. Numer. Math. 44, (1986), 203-220.

[8] _____. A note on a posteriori error bound of Zabreiko and Nguen for Zincenko's iteration. Numer. Funct. Anal. and Optimiz., 9, (9 and 10), (1987), 987-994.

[9] P.P. Zabrejko and D.F. Nguen. The majorant method in the theory of Newton-Kantorovich approximations and the Ptâk error estimates. Numer. Funct. Anal. Optimiz. 9, (1987), 671-684.

[10] A.I. Zincenko. Some approximate methods of solving equations with non-differentiable operators. (Ukrainian). Dopovidi Akad. Navk. Ukrain. RSR (1963), 156-161.

Rev. Academia de Ciencias. Zaragoza **48** *(1993).*

RESULTS ON FIXED POINT THEOREMS SATISFYING A RATIONAL INEQUALITY.

F.U. Rehman*, M.S. Khan** and B. Ahmad***

Abstract.

In this paper, we give fixed point theorems satisfying a rational inequality.

Preliminaries.

Let X denote a complete metric space unless otherwise stated. Define for any non-empty subsets A, B, of X,

$$D(A,B) = \inf\{d(a,b): a \in A,\ b \in B\},$$

$$\delta(A,B) = \sup\{d(a,b): a \in A,\ b \in B\},$$

$$H(A,B) = \max\{\sup\{D(a,B): a \in A\},\ \sup\{D(A,b): b \in B\}\},$$

We denote $CB(X)$ (respt. $BN(X)$), a set of all non-empty closed and bounded (respt. bounded) subsets of X. It is well-known that H is a Hausdorff metric on $CB(X)$. Let A, $B \in CB(X)$ and $k > 1$. Then for each $a \in A$, there is $b \in B$ s.t. $d(a,b) < kH(A,B)$. If $d(A,B) = 0$, then $A = B = \{a\}$ (lemma 1 [4]). In [1], B. Fishe proved theorem 1 with a certain rational inequality. V. Popa [3] gave a similar common fixed point theorem for two multifunctions satisfying a rational inequality and for sequence of multifunctions which generalizes theorem 1 of B. Fisher [1].

In this paper, we prove results of B. Fisher [1] and V. Popa [3] with different rational inequalities and for sequences of self and multivalued mappings. M.S. Khan [2] proved theorem 2, which classifies metric spaces in which each $x \in X$ is a fixed point. We also prove two results which classify metric spaces satisfying a retional inequality (see theorem 7 and 8).

Let $\{S_n\}$ and $\{T_n\}$ be sequences of self maps such that

$$d^q(S_m(x),T_n(y)) \leq \frac{c^q[d^p(y,S_m(x))+d^p(x,T_n(y)]}{d^{p-q}(y,S_m(x))+d^{p-q}(x,T_n(y))}$$

for all x,y in X, for which $d^{p-q}(y,S_m(x)) + d^{p-q}(x,T_n(y)) \neq 0$, $0 < c < 1/2$, $q \geq 1$, $p > q$, $p \geq 2$. Then $\{S_m\}$ and $\{T_n\}$ have a unique common fixed point.

Proof.

Define a sequence $\{x_n\}$ in X as: $x_{2n} \in T_n(x_{2n-1})$, $x_{2n+1} \in S_{n+1}(x_{2n})$ then

$$d^q(x_{2n+1},x_{2n}) = d^q(S_{n+1}(x_{2n}), T_n(x_{2n-1}))$$

$$\leq \frac{c^q[d^p(x_{2n-1},S_{n+1}(x_{2n})) + d^p(x_{2n},T_n(x_{2n-1}))]}{d^{p-q}(x_{2n-1},S_{n+1}(x_{2n})) + d^{p-q}(x_{2n},T_n(x_{2n-1})}$$

$$= \frac{c^q[d^p(x_{2n-1},x_{2n+1}) + d^p(x_{2n},x_{2n}))]}{d^{p-q}(x_{2n-1},x_{2n+1})) + d^{p-q}(x_{2n},x_{2n})} = c^q d^q(x_{2n-1},x_{2n+1})$$

or

$$d^q(x_{2n+1},x_{2n}) \leq c\, d(x_{2n-1},x_{2n+1})$$

or

$$d(x_{2n+1},x_{2n}) \leq a d(x_{2n},x_{2n-1}),$$

where $0 < a = c/1 - c < 1$.

Similarly, $d(x_{2n},x_{2n-1}) \leq a d(x_{2n-1},x_{2n-2})$.

In general $d(x_{n+1},x_n) \leq a d(x_n,x_{n-1}) \leq \ldots\ldots \leq a^n d(x_o,x_1)$

or

$$d(x_{n+1},x_n) \leq a^n d(x_o,x_1).$$

For $m > n$, we have

$$d(x_m,x_n) \leq \sum_{i=0}^{m-n-1} d(x_{n+i},x_{n+i+1}) \leq \sum_{i=0}^{m-n-1} a^{n+i} d(x_o,x_1)$$

$$= d(x_o,x_1)\,[a^n + a^{n+1} + \ldots\ldots + a^{m-1}]$$

$$= a^n d(x_o,x_1)\,[1+a+\ldots.+a^{m-n-1}] \leq a^n d(x_o,x_1)\,[1+a+a^2+\ldots\ldots]$$

or

$$d(x_m,x_n) \leq \frac{a^n}{1-a}\, d(x_0,x_1)$$

When $m,n \to \infty$, $d(x_m,x_n) \to 0$. This shows that $\{x_n\}$ is Cauchy sequence and hence converges to x in X. Then

$$d^q(x_{2n+1},T_m(x)) = d^q(S_{n+1}(x_{2n}),T_m(x)) \leq \frac{c^q[d^p(x,S_{n+1}(x_{2n})) + d^p(x_{2n},T_m(x))]}{d^{p-q}(x,S_{n+1}(x_{2n})) + d^{p-q}(x_{2n},T_m(x))}$$

$$\leq \frac{c^q[d^p(x,x_{2n+1}) + d^p(x_{2n},T_m(x))]}{d^{p-q}(x,x_{2n+1}) + d^{p-q}(x_{2n},T_m(x))}$$

When $n \to \infty$, we have

$$d^q(x\,T_m(x)) \leq c^q d^q(x,T_m(x)) \text{ or } (1-c^q)d^q(x,T_m(x)) \leq 0$$

or $d^q(x,T_m(x)) = 0$ or $d(x,T_m(x)) = 0$ or $x = T_m(x)$, for all m.

Similarly, $x = S_m(x)$, for all m. For the uniqueness, let $y \neq x$, such that $y = T_n(y) = S_n(y)$, for all n. Then

$$d^q(x,y) = d^q(S_m(x),\, T_n(y)) \leq \frac{c^q[d^p(y,S_m(x)) + d^p(x,T_n(y))]}{d^{p-q}(y,S_m(x)) + d^{p-q}(x,T_n(y))}$$

$$= \frac{c^q[d^p(y,x) + d^p(x,y)]}{d^{p-q}(y,x) + d^{p-q}(x,y)} = c^q d^q(x,y)$$

or $(1-c^q)d^q(x,y) \leq 0$ or $d^q(x,y) = 0$ or $d(x,y) = 0$ or $x = y$.

This completes the proof.

Theorem 2.

If $\{S_n\}$ and $\{T_n\}$ are sequences of multivalued mappings from X to $CB(X)$ such that

$$H^q(S_m(x),T_n(y)) \leq \frac{c^{p+1}[D^p(y,S_m(x)) + D^p(x,T_n(y))]}{H^{p-q}(y,S_m(x)) + H^{p-q}(x,T_n(y))}$$

for all x,y in X for which $H^{p-q}(y,S_m(x)) + H^{p-q}(x,T_n(y)) \neq 0$, $0 < c < 1$, $0 < c < 1$, $p > q$, $q \geq 1$, $p \geq 2$, then $\{S_n\}$ and $\{T_n\}$, have a unique common fixed point.

Proof.

Define a sequence $\{x_n\}$ in X as: $x_{2n} \in T_n(x_{2n-1})$, $x_{2n+1} \in S_{2n+1}(x_{2n})$ such that $d(x_{2n+1},x_{2n}) \leq aH(S_{n+1}(x_{2n}),T_n(x_{2n-1}))$, where $a = c^{-1} > 1$. Then

$$d^q(x_{2n+1},x_{2n}) \leq a^q H^q(S_{n+1}(x_{2n}),T_n(x_{2n-1}))$$

$$\leq \frac{a^q c^{p+1}[D^p(x_{2n-1},S_{n+1}(x_{2n})) + D^p(x_{2n},T_n(x_{2n-1}))}{H^{p-q}(x_{2n-1},S_{n+1}(x_{2n})) + H^{p-q}(x_{2n},T_n(x_{2n-1}))}$$

$$\leq \frac{a^q c^{p+1}[d^p(x_{2n-1},x_{2n+1}) + d^p(x_{2n},x_{2n})}{a^{-p+q}[d^{p-q}(x_{2n-1},x_{2n+1})) + d^{p-q}(x_{2n},x_{2n}))}$$

$$= a^p c^{p+1} d^q(x_{2n-1},x_{2n+1})$$

or

$$d(x_{2n+1},x_{2n}) \leq cd(x_{2n-1},x_{2n+1}).$$

Simple calculations give that $\{x_n\}$ is a Cauchy sequence and hence converges to some x in X. To show that $x \in T_m(x)$ for all m, let $u \in T_m(x)$ for all m, then

$$d^q(x_{2n+1},u) \leq a^q H^q(S_{n+1}(x_{2n}),T_m(x)) \leq \frac{a^q c^{p+1}[D^p(x,S_{n+1}(x_{2n})) + D^p(x_{2n},T_m(x))]}{H^{p-q}(x,S_{n+1}(x_{2n})) + H^{p-q}(x_{2n},T_m(x))}$$

$$\leq \frac{c[d^p(x,x_{2n+1}) + d^p(x_{2n},u)]}{d^{p-q}(x,x_{2n+1}) + d^{p-q}(x_{2n},u)}$$

When $n \to \infty$, we have $d^q(x,u) \leq cd^q(x,u)$ or $(1-c)d^q(x,u) \leq 0$ or $d^q(x,u) = 0$ or $d(x,u) = 0$ or $x = u = T_m(x)$. Similarly $x \in S_m(x)$ for all m. The uniqueness can be obtained easily. This completes the proof.

Theorem 3.

If $\{S_n\}$ and $\{T_n\}$ are sequence of multivalued mappings on X to $BN(X)$ such that

$$\delta^q(S_m(x),T_n(y)) \leq \frac{c^{p+q}[H^p(y,S_m(x)) + H^p(x,T_n(y))]}{\delta^{p-q}(y,S_m(x)) + \delta^{p-q}(x,T_n(y))}$$

for all x,y in X, for which $d^{p-q}(y,S_m(x)) + d^{p-q}(x,T_n(y)) \pi 0$, $0 < c < 1/2$, $q \geq 1$, $p \geq 2$, $p > q$, then $\{S_m\}$ and $\{T_n\}$ have a unique common fixed point.

Proof.

Define $\{f_n\}$, $\{g_n\} : X \to X$ such that for each x,y in X , $f_n(x) \in S_n(x)$, $g_n(y) \in T_n(y)$, in such a way that $d(y,fn(x)) \geq cH(y,S_n(x))$ and $d(x,g_n(x)) \geq cH(x,T_n(y))$. Then

$$d^q(f_m(x),g_n(y)) \leq \delta^q(S_m(x), T_n(y)) \leq \frac{c^{p+q}[H^p(y,S_m(x)) + H^p(x,T_n(y))]}{\delta^{p-q}(y,S_m(x)) + \delta^{p-q}(x,T_n(y))}$$

$$\leq \frac{c^{p+q}c^{-p}[d^p(y,f_m(x)) + d^p(x,g_n(y))]}{d^{p-q}(y,f_m(x)) + d^{p-q}(x,g_n(y))}$$

or

$$d^q(f_m(x),g_n(y)) \leq \frac{c^q[d^p(y,f_m(x)) + d^p(x,g_n(y))]}{d^{p-q}(y,f_m(x)) + d^{p-q}(x,g_n(y))} .$$

Then by theorem 1, $\{f_n\}$ and $\{g_n\}$ have a unique common fixed point say u . Then $u = f_n(u) = g_n(u)$. Clearly, $u \in S_n(u)$ and $u \in T_n(u)$ for all n. The uniqueness is easy to established. This completes the proof.

Theorem 4.

Let X be a metric space with respect to two metrics e and d . If X satisfies the conditions:

(1) $e(x,y) \leq d(x,y)$,

(2) X is complete with respect to e,

(3) $S,\{T_n\}:X \to X$ and S is continuous,

(4) $$d^q(Sx,T_n(y)) \leq \frac{c^q\{d^p(y,Sx) + d^p(x,T_n(y))\}}{d^{p-q}(y,Sx) + d^{p-q}(x,T_n(y))}$$

for all x,y in X for which $d^{p-q}(y,Sx) + d^{p-q}(x,T_n(y)) \neq 0$, $0 < c < 1/2$, $p \geq 2$, $q \geq 1$, $p > q$, then S and $\{T_n\}$ have a unique common fixed point.

Proof.

Define a sequence $\{x_n\}$ in X as: $x_{2n} = T_n(x_{2n-1})$ and $x_{2n+1} = Sx_{2n}$. Then as in theorem 1, $\{x_n\}$ is a Cauchy sequence with respect to d

Then by theorem 4, f , g_n , have a unique common fixed point say x in X. Then clearly $x \in Sx$, $x \in T_n(x)$, for all n. The uniqueness can be easily established. This completes the proof.

Theorem 6.

Let $S: X \to X$ be a self mapping satisfying for all x,y in X , $p \geq 2$, $p > q$, $q \geq 1$, the inequality

$$d^q(Sx,Sy) \geq \frac{d^p(x,Sx) + d^p(y,Sy)}{d^{p-q}(x,Sx) + d^{p-q}(y,Sx)}$$

where $d^{p-q}(x,Sx) + d^{p-q}(y,Sx) \neq 0$. Then each x in X is a fixed point of S.

Proof.

Let $x \in X$ be arbitrary. Then

$$0 = d^q(Sx,Sx) \geq \frac{d^p(x,Sx) + d^p(x,Sx)}{d^{p-q}(x,Sx) + d^{p-q}(x,Sx)} = d^q(x,Sx)$$

or $d^q(x,Sx) = 0$, or $d(x,Sx) = 0$, or $x = Sx$. This completes the proof.

Similarly, we can prove

Theorem 7.

Let S be a self map on X s.t. for all x,y in X and $p \geq 2$ $p > q$, $q \geq 1$,

$$d^q(Sx,Sy) \geq \frac{d^p(y,Sx) + d^p(x,Sy)}{d^{p-q}(y,Sx) + d^{p-q}(x,Sy)}$$

where $d^{p-q}(y,Sx) + d^{p-q}(x,Sx) \neq 0$. Then each $x \in X$ is a fixed point of S.

References

[1] B. Fisher: Common fixed points and constant mappings satisfying a rational inequality. Math. Seminar Notes 6, (1978), 29-35.

[2] M.S. Khan: On fixed point theorems. Math. Japonica, 23, Num. 2, (1978), 201-204.

[3] V. Popa: Common fixed points for multifunctions satisfying a rational inequality. Kobe J. of Math. 2 (1985), 23-28.

[4] I.A. Rus: Fixed point theorems for multivalued mappings in complete metric spaces. Math. Japonica, 20 (1975), 21-24.

---

* Government College Kabirwala, Khanewal, Pakistan. Current Address: Department of Mathematics, McGill University, Montreal, Canada.

** P.O. Box 32486, Muscat, Sultanate of Oman.

*** Centre for Advanced Studies in Pure and Applied Mathematics, Bahauddin Zakariya University, Multan, Pakistan.

Rev. Academia de Ciencias. Zaragoza 48 (1993)

ON COMMON FIXED POINTS OF WEAKLY COMMUTING MAPPINGS.

Y.J. Cho, K,S, Park *, T. Mumtaz **, M.S. Khan ***

Abstract.

In this paper, we give common fixed point theorems in metric spaces and 2-metric spaces in terms of an asymptotically regular sequence. Our main theorems generaliza the results of B.M.L. Tiwari - S.L. Singh and M.S. Khan - M. Swa - leh.

I. Introduction and preliminaries.

In [27] S.L. Singh proved the following:

Theorem: *Let (x,d) be a complete metric space. Let S and T be mappings from X into itself such that*

(1.1) T is continuous,

(1.2) $S(X) \subset T(X)$

(1.3) {S,T} is a commuting pair (i.e., $STx = TSx$ *for every* $x \in X$*),*

(1.4) $d(Sx,Sy) \leq a[d(Sx,Tx) + d(Sy,Ty)] + b[d(Sx,Ty) + d(Sy,Tx)] + cd(Tx,Ty)$ *for every* $x,y \in X$*, where a, b, c, are non-negative real numbers such that* $0 < 2a + 2b + c < 1$

then S and T have a unique common fixed point in X.

We remark that if T is the identity mapping on X, then the condition (1.4) reduces to the well-known contraction condition introduced by G.E. Hardy and T.D. Rogers ([10]).

Recently, B.M.L. Tiwari and S.L. Singh ([33]) obtained some common fixed point theorems for a family of mappings satisfying the condition (1.4), which extend and unify several well-known fixed and common fixed point theorems. In [3], H.W. Engl introduced initially the concept of an asymptotically regular sequence in metric spaces, which is very helpful in the study of problems related to the fixed point theory and, using this concept, M.S. Khan - M. Swa - leh ([7]) and B. Ram ([21]) established some fixed point theorems in 2-metric spaces.

On the other hand, the concept of 2-metric spaces has been investigated by S. Gähler in a series of papers [6] - [9] and has been developed extensively by C. Diminnie, R. Freese, K. Iséki, M. Newton, A. White and many othersd ([1]-[2], [4]-[9], [11]-[13], [20]).

A 2-metric space is a set X with a real-valued function d on X x X x X satisfying the following conditions:

(1) For two distinct points x,y in X, there is a point z in X such that $d(x,y,z) \neq 0$,

(2) $d(x,y,z) = 0$ if at least two of x,y,z, are equal,

(3) $d(x,y,z) = d(x,z,y) = d(y,z,x)$,

(4) $d(x,y,z) \leq d(x,y,u) + d(x,u,z) + d(u,y,z)$.

d is called a 2-metric for the space X and (X,d) is called a 2-metric space. It has been shown by S. Gähler ([6]) that although d is a continuous function of any one of its three arguments, it need not be continuous in two arguments. If it is continuous in two arguments, then it is continuous in all three arguments. A d which is continuous in all of its arguments will be called continuous.

K. Iséki ([10]-[13]), B.E. Rhoades ([22]), M.S. Khan ([16]-[17]), S.L. Singh ([28]-[32]) and a number of other mathematicians ([18], [19], [21], [23], [26]) have studied the aspects of fixed point theory in the setting of the 2-metric spaces. They have been motivatied by various concepts known for the metric spaces and have thus introduced analogues of various concepts in the frame work of 2-metric spaces.

In this paper, we give some common fixed point theorems in metric spaces and 2-metric spaces by using an asymptotically regular sequence, which extend and generalize some results of M.S. Khan - M. Swa 1eh ([17]) and B.M.L. Tiwari - S.L. Singh ([33]).

II. Common fixed point theorems in metric spaces.

Throughout this section, let (X,d) be a metric space.

Definition 2.1. Let S and T be mappings from X into itself. A sequence $\{x_n\}$ in X said to be asymptotically S-regular with respect to T if

$$\lim_{n \to \infty} d(Sx_n, Tx_n) = 0$$

If T is the identity mapping on X, then Def. 2.1 reduces to that of H.W. Engl ([3]).

Theorem 2.1. Let (X,d) be a complete metric space. Let $\{S_n\}$ be a family of mappings from X into itself and T be a continuous self-mapping on X such that for each $n \in N$,

(2.1) $\{S_n, T\}$ is a commuting pair,

(2.2) There exists an asymptotically S_n-regular sequence in X with respect to T,

(2.3) $d(S_ix,S_iy) \le a_1d(S_ix,Tx) + a_2d(S_iy,Ty) + a_3d(S_ix,Ty) + a_4d(S_jy,Tx) + a_5d(Tx,Ty)$,

for every $x,y \in X$ and $i,j \in N$, $i \neq j$, where a_i $(i = 1,2,3,4,5)$ are non-negative real number such that $\max\{a_1 + a_2 , a_3 + a_4 + a_5\} < 1$

Then, for each $n \in N$, S_n and T have a unique common fixed point in X.

Proof. Suppose that $\{x_n\}$ is an asymptotically S_i-regular sequence with respecto to T for each $i \in N$. Then $\lim d(Ax,S_ix_n) = 0$ for $n \to \infty$ and for each $i \in N$. Now for each $i,j \in N$, $i \neq j$, consider the inequality:

$$\begin{aligned} d(Tx_n,Tx_m) &\le d(Tx_n,S_ix_n) + d(S_ix_n,S_jx_m) + d(S_jx_m,Tx_m) \\ &\le d(Tx_n,S_ix_n) + \{a_1d(S_ix_n,Tx_n) + a_2d(S_jx_m,Tx_m) + a_3d(S_ix_n,Tx_m) + \\ &\quad + a_4\, d(S_jx_m,Tx_n) + a_5\, d(Tx_n,Tx_m)\} + d(S_jx_m,Tx_m) \end{aligned}$$

Therefore we get

$$d(Tx_n,Tx_m) \le \frac{1+a_1+a_3}{1-a_3-a_4-a_5}\, d(Tx_n,S_ix_n) + \frac{1+a_2+a_4}{1-a_3-a_4-a_5}\, d(Tx_m,S_jx_m)$$

As $m,n \to \infty$, it follows that $\{Tx_n\}$ is a Cauchy sequence in X. Then since (X,d) is complete, $Tx_n \to z \in X$. Also the asymptotic regularity of $\{x_n\}$ would mean that $S_ix_n \to z$. Further the continuity of mapping T yields that $T^2x_n \to Tz$ and $TS_ix_n \to Tz$ for each $i \in N$. On the other hand, since S_i and T are commuting for each $i \in N$, $S_iTx_n \to Tz$. Now for each $i,j \in N$, $i \neq j$, we get

$$\begin{aligned} d(Tz,S_iz) &\le d(Tz,S_jTx_n) + d(S_jTx_n,S_iz) \\ &\le d(Tz,S_jTx_n) + a_1d(S_jTx_n,T^2x_n) + a_2d(S_iz,Tx) + a_3d(S_jTx_n,Tz) + \\ &\quad + a_4d(S_iz,T^2x_n) + a_5d(T^2x_n,Tz) \end{aligned}$$

Letting n tending to infinity, we obtain

$d(Tz,S_iz) \le (a_2+a_4)d(Tz,S_iz)$ for each $i \in N$.

Thus $Tz = S_iz$ for each $i \in N$. Again for each $i,j \in N$, $i \neq j$, we have

$$\begin{aligned} d(S_iTx_n,S_jx_n) &\le a_1d(S_iTx_n,T^2x_n) + a_1d(S_jx_n,Tx_n) + a_1d(S_iTx_n,Tx_n) + \\ &\quad + a_1d(S_jx_n,T^2x_n) + a_1d(T^2x_n,Tx_n) \end{aligned}$$

As $n \to \infty$ we are left with $d(Tz,z) \le (a_3-a_4-a_5)d(Tz,z)$, which implies that $Tz = z$.

Therefore we have $z = Tz = S_iz$ for each $i \in N$. So z is a common fixed point of S_n and T for each $n \in N$. The uniquence of the common fixed point z can be easily established. This complete the proof.

Remark (1): Theorem 2.1 can be regarded as an improvement of Theorem 1 proved by B.M.L. Tiwari and S.L. Singh ([33]). It should be noted that we have not required the condition $S_n(X) \subset A(X)$ in the proof of Theorem 2.1.

Remark (2): We can obtain a multitude of results by choosing S_n and T suitably as shown by B.M.L. Tiwari and S.L. Singh ([33]). Thus our result generalizes several significant fixed and common fixed point theorems.

We also prove the following result on common fixed points under a different set of conditions.

Theorem 2.2. Let (X,d) be a complete metric space. Let $\{S_n\}$ be a family of mappings of X into itself and T be a continuous self-mapping on X such that for each $n \in N$,

(2.5) $S_n(X) \subset T(X)$

(2.6) $\{S_n,T\}$ is a commuting pair,

(2.7) $d(S_nx,S_{n+1}y) \leq a_1\{d(S_nx,Tx) + d(S_{n+1}y,Ty)\} + a_2\{d(S_nx,Ty) + d(S_{n+1}y,Tx)\} + a_3d(Tx,Ty)$,

for all $x,y \in X$, where a_i $(i = 1,2,3)$ are non-negative real numbers such that $\max\{a_1 + a_2, a_3 + a_4 + a_5\} < 1$.

Then there exists a mapping S from X into itself such that T and S have a unique common fixed point in X.

Proof. Let $x_0 \in X$ be arbitrary. Then Tx_0 is also a point of X. Since $S_n(X)$ $T(X)$, we can put $Tx_n = S_n(x_{n-1})$ for each $n \in N$. Then as in the proof of Theorem 1 of B.M.L. Tiwari and S.L. Singh ([33]), we obtain

$$d(Tx_n,Tx_{n+1}) \leq \frac{a_1 + a_2 + a_3}{1 - a_1 - a_2} d(Tx_{n-1},Tx_n)$$

which means that $\{Tx_n\}$ is a Cauchy sequence in X. Hence, bay the completeness of X, $\{Tx_n\}$ converges to some point of X and so $\{S_n x_{n-1}\}$ also converges. Put $Sx = \lim_{n \to \infty} S_nx$

Then we have

$$d(Sx,Sy) = d(\lim_{n \to \infty} S_nx , \lim_{n \to \infty} S_{n+1}y) = \lim_{n \to \infty} (S_nx , S_{n+1}y) =$$

$$\leq \lim_{n \to \infty} [\, a_1\{d(S_nx,Tx) + d(S_{n+1}y,Ty)\} + a_2\{d(S_nx,Ty) + d(S_{n+1}y,Tx)\} + a_3d(Tx,Ty)$$

which implies that

$$d(Sx,Sy) \leq a_1\{d(Sx,Tx) + d(Sy,Ty)\} + a_2\{d(Sy,Ty) + d(Sx,Ty)\} + a_3d(Tx,Ty)$$

Hence by the common fixed point theorem of S.L. Singh ([27]) mentioned in the introduction it follows that T and S have a unique fixed point in X. This completes the proof.

III. Common fixed point theorems in 2-metric spaces.

In this section, we gives a common fixed point theorems in 2-metric space by using the concept of an asymptotically regular sequence. For the following definitions, we refer to [12] - [18], [21] - [26] and [28] - [32].

Definition 3.1: A sequence $\{x_n\}$ in a 2-metric space (X,d) is said to be convergent to a point x in X if $\lim_{n \to \infty} d(x_n,x,a) = 0$ for all a in X. Then x is the limit of the sequence $\{x_n\}$ in X.

Definition 3.2: A sequence $\{x_n\}$ in a 2-metric space (X,d) is said to be a Cauchy sequence if $\lim_{m, n \to \infty} d(x_m,x_n,a) = 0$ for all a in X.

Definition 3.3: A 2-metric space (X,d) in which every Cauchy sequence is convergent is called complete.

Definition 3.4: Let (X,d) be a 2-metric space and T be mapping from X into itself. Then a sequence $\{x_n\}$ in X is said to be asymptotically T-regular if $\lim_{n \to \infty} d(Tx_n,x_n,a) = 0$ for all a in X

Definition 3.5: Let (X,d) be a 2-metric space, and S and T be mapping from X into itself. Then a sequence $\{x_n\}$ in X is said to be asymptotically T-regular with respect to S if for $\lim_{n\to\infty} d(Tx_n,Sx_n,a) = 0$ for all a in X.

If mapping S from X to itself is an identity mapping in Definition 3.5, we obtain Definition 3.4.

Definition 3.6: A mapping T from a 2-metric space (X,d) into itself is said to be sequentially continuous at x if for every sequence $\{x_n\}$ in X such that $\lim_{n\to\infty} d(x_n,x,a) = 0$ for all a in X, $\lim_{n\to\infty} d(Tx_n,Tx,a) = 0$.

Definition 3.7: Let (X,d) be a 2-metric space, and S and T be mappings from X into itself. Then {S,T} is said to be a weakly commuting pair if $d(STx,TSx,a) \leq d(Tx,Sx,a)$ for all a,x in X.

Note that a commuting pair {S,T} in a 2-metric apce (X,d) is weakly commuting, but the converse is not true ([15]).

Now, we give a main theorem.

Theorem 3.1: Let (X,d) be a complete 2-metric space, d continuous an A, S and T be mappings from X into itself such that

(3.1) S and T are sequentially continuous,

(3.2) {A,S} and {A,T} are weakly commuting pairs,

(3.3) There exists an asymptotically A-regular sequence with respect to both S and T,

(3.4) $d(Ax,Ay,a) \le a_1d(Sx,Ax,a) + a_2d(Tx,Ax,a) + a_3d(Sy,Ay,a) + a_4d(Ty,Ay,a)$
$+ a_5d(Sx,Ay,a) + a_6d(Tx,Ay,a) + a_7d(Sy,Ax,a) + a_8d(Ty,Ax,a)$
$+ a_9d(Sx,Ty,a) + a_{10}d(Sy,Tx,a)$

for all x,y,a in X, where a_i (i = 1,2,....,10) is a non-negative real numbers such that $\max\{a_5+a_6+...+a_{10}, a_2+a_3+a_5+a_8+a_9+a_{10}, a_3+a_4+a_5+a_6, a_1+a_2+a_7+a_8\} < 1$

Then A, S and T have a common unique fixed point in X.

Proof.: Let $\{x_n\}$ be an asymptotically A-regular sequence with respect to both S and T. Then, by (3.4), we have

$$d(Ax_n,Ax_m,a) \le a_1d(Sx_n,Ax_n,a) + a_2d(Tx_n,Ax_n,a) + a_3d(Sx_m,Ax_m,a) +$$
$$+ a_4d(Tx_m,Ax_m,a) + a_5d(Sx_n,Ax_m,a) + a_6d(Tx_n,Ax_m,a) +$$
$$+ a_7d(Sx_m,Ax_n,a) + a_8(Tx_m,Ax_n,a) + a_9d(Sx_n,Tx_m,a) +$$
$$+ a_{10}d(Sx_m,Tx_n,a)$$

for all a in X and hence, by condition (4) of 2-metric,

$$(1-a_5-a_6-a_7-a_8-a_9-a_{10})d(Ax_n,Ax_m,a) \le (a_1+a_5+a_9)d(Sx_n,Ax_n,a) + (a_2+a_6+a_{10})d(Tx_n,Ax_n,a) +$$
$$+ (a_3+a_7+a_{10})d(Sx_m,Ax_m,a) + (a_4+a_8+a_9)d(Tx_m,Ax_m,a) +$$
$$+ (a_8+a_9)d(Tx_m,Ax_n,Ax_m) + (a_6+a_{10})d(Ax_m,Ax_n,Tx_m) +$$
$$+ a_5d(Sx_n,Ax_m,Ax_n) + a_7d(Sx_m,Ax_n,Ax_m) +$$
$$+ a_9d(Sx_n.Tx_m,Ax_n) + a_{10}d(Sx_m,Tx_n,Ax_m)$$

for all a in X.

Since $\{x_n\}$ is an asymptotically A-regular sequence with respect to both S and T, as $m,n \to \infty$, we have $(1-a_5-a_6-a_7-a_8-a_9-a_{10})d(Ax_n,Ax_m,a) = 0$ for all a in X. Therefore, $\{Ax_n\}$ is a Cauchy sequence in X. Since (X,d) is a complete 2-metric space, $\{Ax_n\}$ has the limit in X. Call it z, that is, $\lim d(Ax_n,z,a) = 0$ for $n \to \infty$, and for all a in X. Since

$d(Sx_n,z,a) \le d(Sx_n,z,Ax_n) + d(Sx_n,Ax_n,a) + d(Ax_n,z,a) \to 0$ as $n \to \infty$,

$Sx_n \to z$ as . --. Similarly, we have $Tx_n \to z$ as $n \to \infty$. Since S and T are sequentially continuous, it follows that

$SAx_n \to Sz$, $S^2x_n \to Sz$, $STx_n \to Sz$, $TAx_n \to Tz$, $T^2x_n \to Tz$, $TSx_n \to Tz$ as $n \to \infty$. Since {A,T} is a weakly commuting pair,

$$d(ATx_n,Tz,a) \le d(ATx_n,Tz,TAx_n) + d(Tx_n,Ax_n,a) + d(TAx_n,Tz,a)$$

for all a in X. Therefore, we have $ASx_n \to Tz$ as $n \to \infty$. Similarly, we have $ASx_n \to Sz$ as $n \to \infty$. Hence, by (3.4), we have

$$\begin{aligned} d(ASx_n,ATx_n,a) \le\ & a_1d(S^2x_n,ASx_n,a) + a_2d(TSx_n,ASx_n,a) + a_3d(STx_n,ATx_n,a) + \\ & + a_4d(T^2x_n,ATx_n,a) + a_5d(S^2x_n,ATx_n,a) + a_6d(TSx_n,ATx_n,a) + \\ & + a_7d(STx_n,ASx_n,a) + a_8d(T^2x_n,ASx_n,a) + a_9d(S^2x_n,T^2x_n,a) + \\ & + a_{10}d(STx_n,TSx_n,a) . \end{aligned}$$

Since d is continuous, as $n \to \infty$,

$$d(Sz,Tz,a) \le (a_2 + a_3 + a_5 + a_8 + a_9 + a_{10})d(Sz,Tz,a) ,$$

so that, $Sz = Tz$.

Again, by (3.4), for all a in X,

$$\begin{aligned} d(ATx_n,Az,a) \le\ & a_1d(STx_n,ATx_n,a) + a_2d(T^2x_n,ATx_n,a) + a_3d(Sz,Az,a) + \\ & + a_4d(Tz,Az,a) + a_5d(STx_n,Az,a) + a_6d(T^2x_n,Az,a) + \\ & + a_7d(Sz,ATx_n,a) + a_8d(Tz,ATx_n,a) + a_9d(STx_n,Tz,a) + \\ & + a_{10}d(Sz,T^2x_n,a) \\ \le\ & a_1d(STx_n,ATx_n,a) + a_2d(T^2x_n,ATx_n,a) + a_3d(Sz,Az,Tz) + \\ & + a_3d(Sz,Tz,a) + a_3d(Tz,Az,a) + a_4d(Tz,Az,a) + \\ & + a_5d(STx_n,Az,a) + a_6d(T^2x_n,Az,a) + a_7d(Sz,ATx_n,a) + \\ & + a_8d(Tz,STx_n,a) + a_9d(STx_n,Tz,a) + a_{10}d(Sz,T^2x_n,a) \end{aligned}$$

for all a in X.

Since d is continuous for all a in X,

$$d(Tz,Az,a) \le (a_3+a_5)d(Tz,Az,a) + (a_4+a_6)d(Tz,Az,a) = (a_3+a_4+a_5+a_6)d(Tz,Az,a)$$

That is, since $a_3 + a_4 + a_5 + a_6 < 1$, $d(Tz,Az,a) = 0$, so that $Tz = Az$.

Again, by (3.4), for all a in X,

$$d(Az,A^2z,a) \le a_1d(Sz,Az,a) + a_2d(Tz,Az,a) + a_3d(SAz,A^2z,a) + a_4d(TAz,A^2z,a) +$$
$$+ a_5d(Sz,A^2z,a) + a_6d(Tz,A^2z,a) + a_7d(SAz,Az,a) + a_8d(TAz,Az,a) +$$
$$+ a_9d(Sz,TAz,a) + a_{10}d(SAz,Tz,a) = (a_5+a_6+a_7+a_8+a_9+a_{10})d(Az,A^2z,a)$$

Since $a_5 + a_6 + a_7 + a_8 + a_9 + a_{10} < 1$, $d(Az,A^2z,a) = 0$, that is, $Az = A^2z$.

Putting $p = Az$, we have $p = Ap$ and

$$d(Sp,p,a) = d(SAz,Az,a) \le d(SAz,Az,ASz) + d(SAz,ASz,a) + d(ASz,Az,a),$$

that is $d(Sp,p,a) = 0$, so that $Sp = p$. Similarly, $Tp = p$. Thus, p is a common fixed point of A, S and Y.

Next, to prove uniqueness of the common fixed point p, suppose that p and q are common fixed points of A, S and T. Then we have

$$d(p,q,a) = d(Ap,Aq,a) \le a_1d(Sp,Ap,a) + a_2d(Tp,Ap,a) + a_3d(Sq,Aq,a) +$$
$$+ a_4d(Tq,Aq,a) + a_5d(Sp,Aq,a) + a_6d(Tp,Aq,a) +$$
$$+ a_7d(Sq,Ap,a) + a_8d(Tq,Ap,a) + a_9d(Sp,Tq,a) +$$
$$+ a_{10}d(Sp,Tp,a) = (a_5+a_6+a_7+a_8+a_9+a_{10})d(p,q,a).$$

Since $a_5 + a_6 + a_7 + a_8 + a_9 + a_{10} < 1$, $d(p,q,a) = 0$, that is, $p = q$.

Therefore, A, S and T have a unique common fixed point in X.

Remark. In Theorem 3.1, A is sequentially continuous at p if $a_1+a_2+a_7+a_8 < 1$, where p is common fixed point of A, S and T. In fact, let $\{y_n\}$ be any sequence in X with $\lim d(y,p,a) = 0$, for $n \to \infty$, for all a in X, and for $a_1 + a_2 + a_7 + a_8 < 1$.

By (3.4), we have, for all a in X,

$$d(Ay_n,Ap,a) \le a_1d(Sy_n,Ay_n,a) + a_2d(Ty_n,Ay_n,a) + a_3d(Sp,Ap,a) +$$
$$+ a_4d(Tp,Ap,a) + a_5d(Sy_n,Ap,a) + a_6d(Ty_n,Ap,a) +$$
$$+ a_7d(Sp,Ay_n,a) + a_8d(Tp,Ay_n,a) + a_9d(Sy_n,Tp,a) +$$
$$+ a_{10}d(Sp,Ty_n,a)$$
$$\le a_1d(Sy_n,Ay_n,a) + a_1d(Sy_n,Ap,a) + a_1d(Ap,Ay_n,a) +$$
$$+ a_2d(Ty_n,Ay_n,Ap) + a_2d(Ty_n,Ap,a) + a_2d(Ap,Ay_n,a) +$$
$$+ a_5d(Sy_n,Sp,a) + a_6d(Ty_n,Tp,a) + a_7d(Ap,Ay_n,a) +$$
$$+ a_8d(Ap,Ay_n,a) + a_9d(Sy_n,Sp,a) + a_{10}d(Tp,Ty_n,a).$$

Since S and T are sequentially continuous at p,

$$d(Ay_n,Ap,a) \leq (a_1 + a_2 + a_7 + a_8)d(Ap,Ay_n,a)$$

Therefore, we have $\lim d(Ay_n,Ap,a) = 0$, for $n \to \infty$, and for a in X, that is, A is sequentially continuous at p.

As an immediate consequence of Theorem 3.1, we have the following

Corollary 3.2. Let (X,d) be a complete 2-metric space, d continuous and A be a mapping from X into itself satisfying the following condition ([16]):

$$d(Ax,Ay,a) \leq b_1d(x,Ax,a) + b_2d(y,Ay,a) + b_3d(x,Ay,a) + b_4d(y,Ax,a) + b_5d(x,y,a)$$

for all x,y and a in X, where b_i $(i = 1,2,3,4,5)$ is a non-negative real number such that $\max\{b_2 + b_3, b_3 + b_4 + b_5\} < 1$.

If there exists an asymptotically A-regular sequence in X, then A has a unique fixed point in X.

References

1. C. Diminnie and A. White: Non-expansive mappings in linear 2-normed spaces. Math. Japonica, 21 (1976), 127-200.

2. C. Diminnie and A. White: Some geometric remarks concerning strictly 2-convex 2-normed spaces. Mth. Seminar Notes, Kobe Univ. 6 (1978), 245-253.

3. H.W. Engl: Weak convergence of asymptotically regular sequences for non-expansive mappings and connections with certain Chebyshef-centers. Non-linear Analysis, TMA, vol. 1, Nº 5 (1977), 495-501.

4. R. Freese: A 2-metric characterization of Euclidean plane. Math. Ann. 206 (1973), 285-294

5. R. Freese and E. Andalafte: A characterization of 2-betweeness in 2-metric spaces. Canad. J. Math, 18 (1966), 963-968.

6. S. Gähler: 2-metrische Raume und ihre topologisch Straktur. Math. Nachr., 26 (1964), 115-148.

7. S. Gähler: Linear 2-normierte Raume. Math Nachr., 28 (1965), 1-43.

8. S. Gähler: Zur Geometrie 2-metrischer Raume. Rev. Roumaine Math. Pures Appl., 11 (1966), 665-667.

8. S. Gähler: Zur Geometrie 2-metrischer Raume. Rev. Roumaine Math. Pures Appl., 11 (1966), 665-667.

9 S. Gähler: Uber 2-Banach Raume. Math. Nachr., 42 (1969), 335-347.

10. G.E. Hardy and T.D. Rogers: A generalization of a fixed point theorem of Reich. Canad. Math. Bull., 16 (1973), 201-206.

11. K. Iśeki: On non-expansive mappings in strictly convex linear 2-normed spaces. Math Seminar Notes, Kobe Univ., 3 (1975), 125-129.

12. K. Iśeki: A property of orbitally continuous mappings on 2-metric spaces. Math. Seminar Notes, Kobe Univ., 3 (1975), 131-132.

13. K. Iśeki: Fixed point theorems in 2-metric spaces. Math. Seminar Notes, Kobe Univ., 3 (1975), 133-136.

14. K. Iśeki, P.L. Sharma and B.K. Sharma: Contraction type mappings on 2-metric spaces. Math. Japonica, 21 (1976), 67-70.

15. M.D. Khan: A study of fixed point theorems. Thesis. Aligarh Muslim Univ. (1984).

16. M.D. Khan: On convergence of sequence of fixed points in 2-metric spaces. Indian J. Pure Appl. Math., 10 (1979), 1062-1067.

17. M.S. Khan and M. Swa-leh: Results concerning fixed points in 2-metric spaces. Math Japonica, 29(4) (1984), 519-525.

18. S.N. Lal and A.K. Singh: An analogue of Banach's contraction principle for 2-metric spaces. Bull. Austral. Math. Soc., 18 (1978), 137-143.

19. S.N. Lal: Invariant points of generalized non-expansive mappings in 2-metric spaces. Indian J. Math., 30 (1978), 71-76.

20. M. Newton: Uniform and strict convexity in linear 2-normed spaces. Doctoral Dissertation. St. Louis Univ. (1979).

21. B. Ram: Existence of fixed points in 2-metric spaces. Ph. D. Thesis. Garhwal Univ. Srinagar (1982).

22. B.E. Rhoades: Contraction type mappings on 2-metric spaces. Math. Nachr., 91 (1979), 151-155.

23. A.K. Sharma: On fixed points in 2-metric spaces. Math. Seminar Notes, Kobe Univ., 6 (1978), 467-473.

24. A.K. Sharma: A study of fixed points of mappings in metric and 2-metric spaces. Math. Seminar Notes, Kobe Univ., 7 (1979), 291-295.

25. A.K. Sharma: A generalization of Banach contraction principle to 2-metric spaces. Math. Seminar Notes, Kobe Univ., 7 (1979), 191-195.

26. A.K. Sharma: A note on fixed points in 2-metric spaces. Indian J. Pure and Appl. Math. 11 (1980), 1580-1583.

27. S.L. Singh: On common fixed points of commuting mappings. Math. Seminar Notes, Kobe Univ., 5 (1977), 131-134.

28. S.L. Singh: Some contraction type principles on 2-metric spaces and applications. Math. Seminar Notes, Kobe Univ., 7 (1979), 1-11.

29. S.L. Singh: A fixed point theorem in a 2-metric space. Math. Edu(Siwan), 14 (1980), A. 53-54.

30. S.L. Singh and B. Ram: A note on the convergence of sequence of mappings and their common fixed points in a 2-metric space. Math. Seminar Notes, Kobe Univ., 9 (1981), 181-185.

31. S.L. Singh and Virencha: Coincidence theorems on 2-metric spaces. IndiAn J. Phys. Nat. Sci., 2(B) (1982), 32-35.

32. S.L. Singh, B.M.L. Tiwari and K.V. Gupta: Common fixed points of commuting mappings in 2-metric spaces and an application. Math. Nachr., 95 (1980), 293-297.

33. B.M.L. Tiwari and S.L. Singh: Common fixed points of mappings in complete metric spaces. Proc. Nat. Acad. Sci. India, 51(A), I, (1981), 41-44.

---

* Department of Mathematics. Gyeongsang National University. Jinju 660-701. KOREA.

** Department of Mathematics. Aligarh Muslim University. Aligarh 202001. INDIA

*** Department of Mathematics & Computing. Sultan Qaboos University. P.O. Box 32486. Alkhod, Muscat. SULTANATE OF OMAN.

Rev. Academia de Ciencias. Zaragoza **48** *(1993)*

A NEW PROOF FOR THE ORTHOGONAL PROPERTY OF LAGUERRE POLYNOMIALS

by

SADHANA MISHRA

V.B.R.I. Polytechnic, Vidya Bhawan Rural Institute,
Udaipur, INDIA.

Abstract

In this note, we present a new method for establishing the orthogonal property of the Laguerre polynomials.

1. Introduction

The Laguerre polinomials constitute an important, and a rather wide class of hypergeometric polynomials with many applications, particularly in mathematical physics. Their orthogonal property is usually derived by the use of the associated differential equation and Rodrigue's formula. In this note, we introduce a direct method of proof, which is much simpler and elegant to establish the orthogonal property of Laguerre polynomials.

2. Orthogonal property

For $\mathrm{Re}(a) > -1$, the orthogonal property of Laguerre polynomials with the weight function $e^{-x} x^{-a}$ is given by:

$$\int_0^\infty e^{-x}\, x^a\, L_m^a(x)\, L_n^a(x)\, dx = \qquad (2.1)$$

Proof : In view of the relation [1, pag. 311,(9,49)]

$$L_m^a(x) = \frac{(a+1)_m}{m!}\, {}_1F_1(-m;a+1;x) ,$$

the left hand side of (2.1) can be written as

$$\frac{(a+1)_m}{m!}\int_0^\infty e^{-x}\, x^{a}{}_1F_1(-m;a+1;x)\, L_n^a(x)\, dx \qquad (2,2)$$

Now expressing the hypergeometric function in the integrand as its series representation [2,pag. 182,(1)], we have

$$\frac{(a+1)_m}{m!}\sum_{r=0}^{\infty}\frac{(-m)_r}{(a+1)_r r!}\int_0^\infty e^{-x}\, x^{a+r}\, L_n^a(x)\, dx$$

On applying [3,pag. 292,(1)]

$$\int_0^\infty e^{-x}\, x^{b-1}\, L_n^a(x)\, dx = \frac{\Gamma(a-b+n+1)\Gamma(b)}{n!(a-b+1)}, \qquad \mathrm{Re}(b) > 0$$

and simplifying, (2.2) reduces to the form

$$\frac{(a+1)_m}{m!n!}\sum_{r=0}^{m}\frac{(-m)_r(-r)_n\Gamma(a+1)}{r!} \qquad (2.3)$$

If $r < n$, the numerator of (2.3) vanishes, and since r runs from 0 to m, it follows that (2.3) also vanishes when $n<m$. Now, it is clear that for $m \neq n$ all terms of (2.3) vanish.

When $m = n$, using the standard resul [1,pag.274]:

$$(-r)_n = \begin{cases} \dfrac{(-1)^n r!}{(r-n)!}, & \text{if } 0 \le n \le r; \\ 0, & \text{if } n > r, \end{cases}$$

and simplifying (2.3), we get

$$\frac{(a+1)_n\Gamma(a+1)}{n!},$$

which yields the orthogonal property (2.1).

References

[1] Andrews, L.C.: *Special Functions for Engineers and Applied Mathematicians*. Macmillan Publishing Co., New York, 1985.

[2] Erdélyi, A. et al: *Higher transcendental functions, vol. 1*. McGraw-Hill, New York , 1953.

[3] Erdélyi, A. et al: *Tables of integral transforms, vol. 2*. McGraw-Hill, New York , 1953.

Rev. Academia de Ciencias. Zaragoza. 48 (1993)

PARTIAL DIFFERENTIAL EQUATIONS OF HOMOTHETICITY

by

Juan Carlos CANDEAL [1] and Esteban INDURAIN [2]

1) Departamento de Análisis Económico. Universidad de Zaragoza. Spain.

2) Departamento de Matemática e Informática. Universidad Pública de Navarra. Pamplona. Spain.

Abstract : We introduce here a system of partial differential equations that characterizes homothetic differentiable functions defined on real cones.

A.M.S. Subject Classification (1991) : Primary : 26 B 35, 35 A 08 Secondary : 26 B 10, 35 F 20 , 90 A .

Key Words : Homothetic functions. Systems of partial differential equations.

1. Introduction and motivation.

"Homothetic function" is a term which refers to some extension of the concept of a homogeneous function. Different definitions of the concept of "homotheticity", *not all of them being equivalent* , can be found in the literature. (See , for instance, Eichhorn [1968,1969] ; Kats [1970] ; Whitaker and Mc. Callum [1971] ; Shephard [1972] ; Aczél and Dhombres [1989] , or Candeal and Induráin [1992 a,b]).

Among the definitions of homotheticity, there exists a *functional equation representation* (Färe [1973]) :

(I) << $f(\lambda x) = \Gamma(\lambda , f(x))$, *for every* $x \in X$ *and* $\lambda \in \mathbb{R}_{++}$, " Γ " *being a two-variables function* >>.

The existence of that representation is equivalent to the *scaling-invariance* of the function "f" :

(II) $<<$ $f(x) = f(y)$ *if and only if* $f(\lambda x) = f(\lambda y)$, *for every* $x, y \in X$, *and* $\lambda \in \mathbb{R}_{++}$ $>>$.

Condition (II) appears in the definition of homothetic functions introduced in Whitaker and Mc. Callum [1971] .

Last definition is, perhaps, the key for applications of the concept of homotheticity in Economics. As an intuitive example , consider an individual declaring that the commodity bundle "x" is , for her, preferred to the commodity bundle "y" . It seems acceptable that she should declare preferred a k-fold multiple of "x" when compared to a k-fold multiple of "y", for any positive scale "k" . (These kind of questions are analyzed , for instance, in Chipman [1974] ; Nadiri [1982] , or Shafer and Sonnenschein [1982]) .

2. Notation and previous concepts.

We shall denote $\mathbb{R}_{++} = \{x \in \mathbb{R} ; x > 0\}$ and $\mathbb{R}^n_{++} = \mathbb{R}_{++} \times \ldots$ (n-times) $\ldots \times \mathbb{R}_{++}$.

In what follows , $X \subset \mathbb{R}^n_{++}$ will be a connected open *real cone* , and $f : X \longrightarrow \mathbb{R}$ a real function.

The function f is said to be :

(i) Homogeneous of degree m ($m \in \mathbb{R}$) if $f(\lambda x) = \lambda^m f(x)$, for every $\lambda \in \mathbb{R}_{++}$ and $x \in X$,

(ii) homothetic (in the sense of Färe [1973]) if it verifies the functional equation (I) (or its equivalent (II)).

3. Partial differential equations of homotheticity.

Functional equations (I) and (II) are , in a way, non-classical. (For *classical* "functional equations" one usually means *differential or integral equations*).

This in mind, it sounds interesting to look for differential equations that characterize the homotheticity , at least for the differentiable case.

In this section, we characterize differentiable homothetic functions as the solution of a system of partial differential equations

Consider the following equation :

$$(III) \qquad \nabla f(x) = H(x) \,.\, S(\nabla f(x)) \quad ,$$

where "S" denotes the shift operator on $\mathbb{R}^n$ defined by $S(x_1,x_2,\ldots,x_n) = (x_2,\ldots,x_n,x_1)$; $f : X \longrightarrow \mathbb{R}$ is a real function of class $\mathcal{C}^1(X)$; $\nabla f(x)$ denotes its gradient (column matrix) at point $x \in X$; and "H" is a homogeneous of zero degree diagonal matrix of real functions, i.e. :

$$H_1(x) = \begin{pmatrix} h_1(x) & 0 & \ldots & 0 \\ 0 & h_2(x) & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & h_n(x) \end{pmatrix}$$

the functions h_i $(i = 1,\ldots,n)$ being homogeneous of zero degree, $h_i(x) \neq 0$ $(i = 1,\ldots,n \; ; \; x \in X)$, and $h_n(x) = 1 / [h_1(x).h_2(x). \ldots .h_n(x)]$.

Equation (III) is equivalent to the following system of partial differential equations :

$$(III^*) \quad \begin{cases} \partial f/\partial x_1 (\lambda x) = h_1(x).\, \partial f/\partial x_2 (\lambda x) \quad , \\ \partial f/\partial x_2 (\lambda x) = h_2(x).\, \partial f/\partial x_3 (\lambda x) \quad , \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad , \\ \partial f/\partial x_{n-1} (\lambda x) = h_{n-1}(x).\, \partial f/\partial x_n (\lambda x) \quad , \\ \partial f/\partial x_n (\lambda x) = h_n(x).\, \partial f/\partial x_1 (\lambda x) \quad , \end{cases}$$

for every $\lambda \in \mathbb{R}_{++}$, $x \in X$.

Henceforward , we shall deal with a function $f : X \longrightarrow \mathbb{R}$ *of class* $\mathcal{C}^1(X)$ *, and such that* $f_i = \partial f/\partial x_i > 0$ $(i = 1, \ldots , n)$.

We are ready to present now our main result.

THEOREM :

The following statements are equivalent :

(i) "f" is a solution of system (III) ,*
(ii) "f" is homothetic .

Proof :

(i) $\Rightarrow$ *(ii)*

Consider a *level-surface* relative to "f", given by $\{x = (s , z) \in X ; s \in \mathbb{R}^{n-1}_{++} , z \in \mathbb{R}_{++} , f(x) = C \}$. Fix a point $x^* = (a,b)$ in that level surface (i.e. : $s \in \mathbb{R}^{n-1}_{++}$, $z \in \mathbb{R}_{++}$).

Since $\partial f/\partial x_n (x^*) > 0$, there exists an implicit function " ζ " defined in a neighbourhood of "a" , $\mathcal{U}$, such that $\zeta(a) = b$, and $f(s, \zeta(s)) = C$, for every $s \in \mathcal{U}$.

Observe that , for every $s \in \mathcal{U}$ and $i \in \{1, \ldots , n-1\}$ it holds :

$$\partial f/\partial x_i (s, \zeta(s)) + \partial f/\partial x_n (s, \zeta(s)) . \partial \zeta/\partial x_i(s) = 0 .$$

Thus $\partial \zeta/\partial x_i(s) = - [(\partial f/\partial x_i (s, \zeta(s))) / (\partial f/\partial x_n (s, \zeta(s)))]$
$= - [h_{i+1}(s, \zeta(s)) . h_{i+2}(s, \zeta(s)) . \ldots . h_{n-1}(s, \zeta(s))]$.

Therefore, by hypotheses of (i) , it follows that $\partial \zeta/\partial x_i(s)$ is homogeneous of degree zero.

Now, let $\lambda \in \mathbb{R}_{++}$. At least in the neighbourhood $\lambda \mathcal{U}$ of $(\lambda a, \lambda b)$, the function " Φ " given by $\Phi(t) = f(t , \lambda . \zeta(t/\lambda))$ is well defined.

To conclude that (i) implies (ii) , it is enough to see that this function " Φ " is constant , or, equivalently, it has zero gradient .

Actually, fixed $i \in \{1, \dots , n-1\}$, we have :

$$\partial\Phi/\partial x_i\,(t) = \partial f/\partial x_i\,(t, \lambda . \zeta(t/\lambda)) + \lambda . \partial f/\partial x_n (t , \lambda . \zeta(t/\lambda)) . (1/\lambda) . \partial\zeta/\partial x_i(t/\lambda) = \partial f/\partial x_i\,(t, \lambda . \zeta(t/\lambda)) + \partial f/\partial x_n (t , \lambda . \zeta(t/\lambda)) . \partial\zeta/\partial x_i(t/\lambda) = \partial f/\partial x_i\,(t, \lambda . \zeta(t/\lambda)) - [(\partial f/\partial x_i\,(t, \lambda . \zeta(t/\lambda))) / (\partial\zeta/\partial x_i(t))] . \partial\zeta/\partial x_i(t/\lambda) = 0 .$$

(ii) $\Rightarrow$ (i)

It is enough to prove that the function $(\partial f/\partial x_i) / (\partial f/\partial x_n) : X \longrightarrow \mathbb{R}$ is homogeneous of degree zero, for every $i \in \{1, \dots , n-1\}$.

Fix $x^* = (x_1^*,\dots,x_n^*) \in X$. Set $A(x^*) = \{z \in X ; f(z) = f(x^*)\}$.

By the Implicit Function Theorem, there exist a neighbourhood "$\mathcal{U}$" of $(x_1^*,\dots,x_{i-1}^*,x_{i+1}^*,\dots,x_n^*)$, and a differentiable function " φ " defined on "$\mathcal{U}$" , such that $\varphi(x_1^*,\dots, x_{i-1}^*, x_{i+1}^*,\dots, x_n^*) = x_i^*$; $(\partial\varphi/\partial x_i)(x_1 ,\dots, x_{i-1}, x_{i+1} ,\dots, x_n) = -[(\partial f/\partial x_i) / (\partial f/\partial x_n)](x_1,\dots,x_n)$.

Given $\lambda \in \mathbb{R}_{++}$, set $A(\lambda.x^*) = \{s \in X ; f(s) = f(\lambda.x^*)\}$. The homotheticity of "f" implies that $A(\lambda.x^*) = \lambda . A(x^*)$.

As above, there exist a neighbourhood "$\mathcal{W}$" of $(\lambda.x_1^*, \dots , \lambda.x_{i-1}^*, \lambda.x_{i+1}^*, \dots , \lambda.x_n^*)$, and a differentiable function " $\varphi\#$ " defined on "$\mathcal{W}$" , such that $\varphi\#(\lambda.x_1^*, \dots , \lambda.x_{i-1}^*, \lambda.x_{i+1}^*, \dots , \lambda.x_n^*) = \lambda . x_i^*$; $(\partial\varphi\#/\partial x_i)(y_1, \dots , y_{i-1}, y_{i+1}, \dots , y_n) = -[(\partial f/\partial x_i) / (\partial f/\partial x_n)](y_1, \dots , y_n)$.

Now, by the *uniqueness* of implicit functions , it follows that on the subset $\mathcal{W} \cap \lambda.\mathcal{U}$ it holds :

$$\varphi\#(y_1,\dots,y_{i-1}, y_{i+1},\dots,y_n) = \lambda.\varphi((y_1/\lambda),\dots,(y_{i-1}/\lambda),(y_{i+1}/\lambda),\dots,(y_n/\lambda)).$$

Hence, we have :

$$- [(\partial f/\partial x_i) / (\partial f/\partial x_n)] (\lambda.x_1, \dots , \lambda.x_{i-1}, \lambda.x_{i+1}, \dots , \lambda.x_n) =$$

$$\partial\varphi\#/\partial x_i (\lambda.x_1, \dots , \lambda.x_{i-1}, \lambda.x_{i+1}, \dots , \lambda.x_n) = (\partial\varphi/\partial x_i) (x_1 ,\dots, x_{i-1}, x_{i+1},\dots, x_n) = - [(\partial f/\partial x_i) / (\partial f/\partial x_n)] (x_1,\dots,x_n) .$$

Therefore $(\partial f/\partial x_i) / (\partial f/\partial x_n)$ is homogeneous of degree zero.

This concludes the proof.

FINAL REMARKS :

(a) In Candeal and Induráin [1992 a,b] we studied the *structure* of homothetic functions and give a complete classification of them. For example, a continuous homothetic function "f" on a connected open cone "X" is either constant (in particular homogeneous of degree zero), or there exist continuous functions $g: X \longrightarrow \mathbb{R}$ and $h : \mathbb{R} \longrightarrow \mathbb{R}$ with "g" homogeneous of degree one, "h" injective , and $f = h \circ g$.

(b) The characterization in Remark (a) corresponds to the most common definition of homotheticity given in the Mathematical Economics literature, namely :

A function $f : X \longrightarrow \mathbb{R}$ is said to be *homothetic in the classical sense of Economics* if there exist

$h : X \longrightarrow \mathbb{R}$ and $g : \mathbb{R} \longrightarrow \mathbb{R}$ such that $f = g \circ h$, "g" being injective and "h" homogeneus of degree one .

(c) Several authors call *"homothetic in the classical sense of Economics"* only the particular case of "f" being differentiable such that $f = g \circ h$, with $g' > 0$, and "h" being homogeneous of degree one. That is the definition found in Shephard [1953] ; Kats [1970] or Silberberg [1990] , p. 97 .

(d) Under the hypotheses of the Theorem, the existence of a decomposition $f = g \circ h$, as that in Remark (c) , (i.e. : $g' > 0$, and "h" being homogeneous of degree one) *characterizes* homotheticity. (i.e. : under differentiability assumptions, homotheticity in the sense of Färe is equivalent to homotheticity in the classical sense of Economics).

To prove this assertion, first observe that being $f = g \circ h$, the partial derivatives of "h" are homogeneous of zero degree. Thus, "f"

verifies a system of equations like (III*). Conversely, suppose that "f" is homothetic in the sense of Färe. Fix an element $e \in X$. Given $x \in X$, there exists a unique $\lambda_x \in \mathbb{R}_{++}$ such that $f(x) = f(\lambda_x . e)$ (to convince you, notice that "f" is continous and increasing in each variable, and use an standard connectedness argument).

Finally, consider the functions $h : X \longrightarrow \mathbb{R}_{++}$, defined by $h(x) = \lambda_x$, and $g : \mathbb{R}_{++} \longrightarrow \mathbb{R}_{++}$ defined by $g(x) = f(\lambda_x . e)$.

(e) In Lau [1969] a characterization of homothetic functions through a system of partial differential equations was also obtained. Nevertheless, the approach in that paper is much less elementary than in our main result. For example, several additional conditions related to partial derivatives of second order, integrability , quasiconcavity, etc. , were considered there.

(f) The case $n = 2$ has an interesting interpretation : Given a function $f(x,y)$ such that $\partial f/\partial x\ (\lambda x, \lambda y) = h_1(x,y).\ \partial f/\partial y\ (\lambda x, \lambda y)$; for every $\lambda \in \mathbb{R}_{++}$, $h_1(x,y)$ being a function homogeneous of zero degree , it follows that the level-curves $f(x,y) = k$ $(k \in \mathbb{R})$ are integral solutions of the homogeneous ordinary differential equation $y' = - h_1(x,y)$. It is well known that the integral-curves , or solutions, of an homogeneous O.D.E. are homothetic (in the geometrical sense) with respect to the origin of the plane. (See, for instance, Valiron [1950] pp. 145 and ff.) .

4. References.

ACZÉL, J. and J. DHOMBRES : "Functional equations in several variables with applications to mathematics, information theory and to the natural and social sciences". Cambridge University Press, U.K. 1989.

CANDEAL, J. C. and E. INDURAIN : "Estudio analítico sobre funciones homotéticas". Revista Española de Economía 9 (1) , 103-114. (1992 a).

CANDEAL, J.C. and E. INDURAIN : "On the structure of homothetic functions". Aequationes Mathematicae. (1992 b , to appear).

CHIPMAN, J. S. : "Homothetic preferences and aggregation". Journal of Economic Theory 8 , 26-38. (1974).

EICHHORN, W. : "Behandlung zweier auf die Untersuchung von Funktionalgleichungen führender produktiontheoretischer Probleme". Jahrbücher für Nationalökonomie und Statistik 181, 334-342. (1968).

EICHHORN, W. : "Eine Verallgemeinerung des Begriffs der homogenen Produktionsfunktion". Unternehmenforschung 13 (2), 99-109. (1969).

FÄRE, R. : "On scaling laws for production functions". Zeitschrift für Operations Research, pp. 195-205. (1973).

KATS, A. : "Comments on the definition of homogeneous and homothetic functions". Journal of Economic Theory 2, 310-313. (1970).

LAU, L. J. : "Duality and the structure of utility functions". Journal of Economic Theory 1, 374-96. (1969).

NADIRI, M. I. : "Producers Theory". Chapter 10 of the "Handbook of Mathematical Economics, vol II". (Edited by K. J. Arrow and M. D. Intrilligator). North Holland . Amsterdam. 1982 .

SHAFER, W. and H. SONNENSCHEIN : "Market demand and excess demand functions" . Chapter 14 of the "Handbook of Mathematical Economics, vol II". (Edited by K. J. Arrow and M. D. Intrilligator). North Holland . Amsterdam. 1982 .

SHEPHARD, R. W. : "Cost and Production Functions". Princeton University Press. Princeton, NJ. , U.S.A. 1953.

SHEPHARD, R. W. : "Comments on homogeneous production functions". Journal of Economic Theory 4 , 101-102. (1972).

SILBERBERG, E. : "The structure of Economics : A mathematical analysis". (Second edition). Mc Graw Hill. New York. 1990.

VALIRON, G. : "Équations fonctionnelles. Applications". Masson. Paris. 1950.

WHITAKER, J. K. and B. T. Mc. CALLUM : "On homotheticity of Production Functions". Western Economic Journal 9, 57-63. (1971).

Rev. Academia de Ciencias. Zaragoza **48** *(1993).*

CORRECCION DE ORBITAS DE PARES VISUALES UTILIZANDO SOLAMENTE DIFERENCIAS (O-C) EN ANGULOS DE POSICION

R. Cid y C. Longás
Departamento de Física Teórica
Sección de Astronomía
Universidad de ZARAGOZA

Abstract.

In this paper, we have developed three methods for obtaining the correction of orbits of double stars, using only differences (O-C) in position angles. The results are employed for the correction of the orbit ADS 15972, making use of two of these methods.

1. Introducción.

El cálculo y corrección de órbitas de órbitas de estrellas dobles visuales se basa en un conjunto de observaciones del tipo (ρ, θ, t) que son efectuadas por numerosos astrónomos durante años, de manera que para cada instante t tomando una de las estrellas como *estrella principal* E_0 (la más brillante o la elegida por el primer observador, si ambas son de la misma magnitud), la *estrella satélite* E es referida a la anterior y se supone que describe una órbita kepleriana llamada *órbita relativa*. La proyección cilíndrica de esta órbita sobre el plano tangente a la esfera celeste en E_0 recibe el nombre de *órbita aparente*. De esta forma, si E' es la proyección de E sobre la órbita aparente, los datos de observación (ρ, θ, t) vienen dados por la distancia $\rho = \text{dist}(E_0E')$ y el ángulo de posición $\theta = NE_0E'$, siendo N la dirección del Norte.

Por otra parte, el desconocimiento de las masas de ambas estrellas obliga a introducir el periodo P como elemento orbital adicional, de forma que en la determinación de órbitas de estrellas dobles suelen emplearse los siguientes *elementos orbitales:*

P = *periodo,* o tiempo en años que emplea la estrella satélite en su trayectoria en torno a la estrella principal, sustituido en ocasiones por el *movimiento medio* $n = 2\pi/P$.

T = *época de paso por el periastro,* o instante t en que la estrella satélite se encuentra sobre el semieje mayor de la órbita relativa, en su posición más próxima a la estrella principal.

a = *semieje mayor* de la órbita relativa

e = *excentricidad* de dicha órbita

I = *inclinación,* o ángulo diedro que forman los planos de ambas órbitas (aparente y relativa) contado de 0º a 90º para órbitas de movimiento directo y de 90º a 180º para las de movimiento retrógrado.

Ω = *ángulo del nodo,* ($\Omega = NE_0E'$), siendo E_0E' la recta de intersección de los dos planos fundamentales. Se cuenta de 0º a 180º.

ω = *argumento del periastro,* que define el ángulo entre el nodo Ω y el periastro T, contado de 0º a 360º.

La relaciones que existen entre los elementos (ρ, θ) de la órbita aparente y los correspondientes (r, f) de la órbita relativa, donde r es el *radio vector* y f *la anomalía verdera,* son las siguientes:

$$\rho \cos(\theta - \Omega) = r \cos(\omega + f) \qquad \rho \operatorname{sen}(\theta - \Omega) = r \operatorname{sen}(\omega + f) \cos I \qquad (1.1)$$

Con estas notaciones, y siendo M la *anomalía media* y E la *anomalía excéntrica*, el cálculo de las efemérides (ρ, θ) correspondientes a una época t, se obtienen por la secuencia de fórmulas

$$M = n(t - T), \qquad E - e \operatorname{sen} E = M, \qquad \operatorname{tag} \frac{f}{2} = \sqrt{\frac{1+e}{1-e}} \operatorname{tag} \frac{E}{2}, \quad (1.2)$$

$$r = a(1 - e\cos E) \qquad \tan(\theta-\Omega) = \tan(\omega+f)\cos I \qquad \rho = r\frac{\cos(\omega+f)}{\cos(\theta-\Omega)} \tag{1.2}$$

Para evitar repeticiones en lo que sigue, es conveniente introducir las constantes de Innes, A, B, F, G, que vienen definidas por las fórmulas:

$$\begin{aligned} A &= a(\cos\Omega\cos\omega - \operatorname{sen}\Omega\operatorname{sen}\omega\cos I) \\ B &= a(\operatorname{sen}\Omega\cos\omega + \cos\Omega\operatorname{sen}\omega\cos I) \\ F &= a(-\cos\Omega\operatorname{sen}\omega - \operatorname{sen}\Omega\cos\omega\cos I) \\ G &= a(-\operatorname{sen}\Omega\operatorname{sen}\omega + \cos\Omega\cos\omega\cos I) \end{aligned} \tag{1.3}$$

Invirtiendo esta fórmulas se llega facilmente a las siguientes:

$$\tan(\omega + \Omega) = (B - F)/(A + G)$$

$$\tan(\omega - \Omega) = -(B + F)/(A - G) \tag{1.4}$$

$$\tan^2\left(\frac{I}{2}\right) = \left(\frac{A-G}{A+G}\right)\frac{\cos(\omega+\Omega)}{\cos(\omega-\Omega)} = -\left(\frac{B+F}{B-F}\right)\frac{\operatorname{sen}(\omega+\Omega)}{\operatorname{sen}(\omega-\Omega)}$$

que permiten calcular los elementos Ω, ω, I, en función de las constantes de Innes, sirviendo la última, que ha de ser necesariamente positiva, para discriminar los verdaderos valores de Ω y ω.

2. Cálculo de órbitas por el método de R. Cid.

Siendo $K = a\sqrt{1-e^2}\cos I$, y por medio de la ecuación de Thiele

$$n(t_k - t_i) - [E_k - E_i - \operatorname{sen}(E_k - E_i)] = \frac{\rho_k\rho_i\operatorname{sen}(\theta_k-\theta_i)}{K}$$

en este método son utilizados tres lugares normales, (ρ_2,θ_2,t_2) , (ρ_3,θ_3,t_3) , (ρ_4,θ_4,t_4) y una observación adicional (θ_1,t_1), que sirven de base a los cálculos, llegándose finalmente a un sistema de tres ecuaciones

$$\begin{aligned} &F(V-U) - pF(V) + qF(U) = 0 \\ &hF(W-V) - kF(W-U) + F(W) = 0 \\ &Y = \Phi(VU) - (1-N)\Phi(WU) + (1-Q)\Phi(WV) = 0 \end{aligned} \tag{2.1}$$

con las incógnitas $U = E_4 - E_3$, $V = E_4 - E_2$, $W = E_4 - E_1$, y donde las funciónes $F(X)$ y $\Phi(XY)$ están definidas en la forma

$$F(X) = X - \text{sen } X \qquad \Phi(XY) = -\text{sen } X + \text{sen } Y + \text{sen}(X-Y) \qquad (2.2)$$

Las cantidades N, Q, R, S, dependen exclusivamente de los datos, por medio de las igualdades:

$$N = \frac{\rho_2 \,\text{sen}(\theta_2-\theta_1)}{\rho_4 \,\text{sen}(\theta_4-\theta_1)} \qquad Q = \frac{\rho_3 \,\text{sen}(\theta_3-\theta_1)}{\rho_4 \,\text{sen}(\theta_4-\theta_1)} \qquad (2.3)_1$$

$$R = \frac{\rho_3 \,\text{sen}(\theta_3-\theta_2)}{\rho_4 \,\text{sen}(\theta_4-\theta_2)} \qquad S = \frac{\rho_2 \,\text{sen}(\theta_3-\theta_2)}{\rho_4 \,\text{sen}(\theta_4-\theta_3)} \qquad (2.3)_2$$

aunque también pueden expresarse, haciendo uso de la ecuación de Thiele, en las formas siguientes

$$N = \frac{n(t_2-t_1) - F(W-V)}{n(t_4-t_1) - F(W)} \qquad Q = \frac{n(t_3-t_1) - F(W-U)}{n(t_4-t_1) - F(W)}$$

$$R = \frac{n(t_3-t_2) - F(V-U)}{n(t_4-t_2) - F(V)} \qquad S = \frac{n(t_3-t_2) - F(V-U)}{n(t_4-t_3) - F(U)}$$

Las cantidades p, q, h, k, dependen de los datos y de las cantidades anteriores, por medio de las igualdades

$$p = \frac{Rt_{23} - RSt_{34}}{Rt_{24} - St_{34}} \qquad q = \frac{St_{23} - RSt_{24}}{Rt_{24} - St_{34}} \qquad (2.4)_1$$

$$h = \frac{t_{13} - Qt_{14}}{Qt_{12} - Nt_{13}} \qquad k = \frac{t_{12} - Nt_{14}}{Qt_{12} - Nt_{13}} \qquad (2.4)_2$$

donde $t_{ik} = t_k - t_i$.

La resolución del sistema (2.1), descrita en el trabajo de R. Cid (1960), puede efectuarse dando valores a V y calculando los correspondientes de U por la primera fórmula y los correspondientes de W por la segunda fórmula, como función de los anteriores. La tercera fórmula sirve de control hasta que se produzca su anulación. Las posibles soluciones del sistema se encuentran entre aquellos valores de V, para los cuales se produce un cambio de signo de Y. La resolución de

dicho sistema puede efectuarse en forma analítica, como fué diseñada en el citado trabajo de R. Cid, o en forma numérica.

Una vez resuelto el sistema (2.1) con respecto a las incógnitas U, V, W, se puede proceder al cálculo de los elementos orbitales del modo siguiente:

a) La definición de las cantidades N, Q, R, S, permite llegar a cuatro ecuaciones, que solamente contienen la incógnita n. Por lo cual pueden obtenerse cuatro valores, cuyo promedio define el movimiento medio y por tanto el periodo.

b) Las ecuaciones de Thiele, aplicadas a los tres pares de índice (2,3), (2,4), (3,4), que no contienen la distancia ρ_1, servirán para el cálculo de la constante K. Por ejemplo

$$K = \frac{\Delta_{24}}{nt_{24} - F(V)}$$

c) Las ecuaciones

$$e \operatorname{sen} E_3 = \frac{R\operatorname{sen}(V-U) - RS\operatorname{sen}U}{RS+R-S}$$

$$e \cos E_3 = \frac{RS\cos U - R\cos(V-U) - S}{RS+R-S}$$

nos proporcionan el valor de la excentricidad e y el verdadero cuadrante de la anomalía excéntrica E_3, de la que se deducen las anomalías

$$E_4 = E_3 + U \ , \qquad E_2 = E_3 - (V-U) \ , \qquad E_1 = E_3 - (W-U)$$

d) Las ecuaciones de Kepler

$$n(t_i - T) = E_i - e \operatorname{sen} E_i \quad (i = 1,2,3,4)$$

nos darán un promedio de la época T de paso por el periastro.

e) Finalmente, para el cálculo de los elementos Ω , ω , I , puede seguirse el proceso dado por R. Cid (Urania, 1960), calculando para tres épocas (t_i, t_j, t_k), [siempre en el orden de la permutación (i,j,k)], los determinantes

$$M_i = \begin{vmatrix} p_j + q_j & 1 - p_j q_j \\ p_k + q_k & 1 - p_k q_k \end{vmatrix} \qquad N_i = \begin{vmatrix} p_j - q_j & 1 + p_j q_j \\ p_k - q_k & 1 + p_k q_k \end{vmatrix}$$

donde $p_i = \text{tag}\,\theta_i$, $q_i = \text{tag}\,f_i$.

Los determinantes M_i , M_j , M_k , N_i , N_j , N_k , nos permitirán calcular los elementos Ω, ω, I, por medio de las igualdades

$$\text{tag}(\omega + \Omega) = -\frac{q_i M_i + q_j M_j + q_k M_k}{M_i + M_j + M_k}$$

$$\text{tag}(\omega - \Omega) = -\frac{q_i N_i + q_j N_j + q_k N_k}{N_i + N_j + N_k}$$

$$\text{tag}^2\left(\frac{I}{2}\right) = \frac{(N_i + N_j + N_k)\cos(\omega + \Omega)}{(M_i + M_j + M_k)\cos(\omega - \Omega)}$$

f) Finalmente, el semieje mayor a se obtendrá por la igualdad

$$a^2 = \frac{K}{\sqrt{1-e^2}\cos I}$$

quedando concluido el cálculo de los elementos orbitales.

3. Métodos de corrección de órbitas que utilizan diferencias observación menos cálculo en ángulos de posición.

Una vez realizado un primer cálculo de la órbita de una estrella doble por cualquier método y calculadas las efemérides correspondientes, se obtienen las diferencias existentes entre los datos de observación (ρ, θ, t) y los calculados por efemérides, que suelen denominarse *diferencias observación menos cálculo* (O-C), en ángulos de posición y distancias, o bien $\Delta\rho$ y $\Delta\theta$.

Con ayuda de estas diferencias puede procederse a una corrección de la órbita calculada, ya sea por el *método de mínimos cuadrados*, en el que se procura hacer mínima la suma de los cuadrados de tales diferencias, o por el *método de mínimos valores absolutos*. en el que se trata de minimizar la suma de valores absolutos de dichas diferencias.

En general, estos métodos están adaptados para el cálculo con ordenadores, de tal manera que la órbita corregida por un primer cálculo puede servir de base a una segunda corrección y así sucesivamente.

Designando por q_i el conjunto de elementos orbitales independientes P , T , e , Ω , ω , I , y por θ_k (k = 1, 2, ..., n) un conjunto de datos de observación, tendremos n relaciones conocidas $\theta_k = f(q_i , t_k)$, que para valores previos $(q_i)_0 \equiv (P_0, T_0, e_0, \Omega_0, \omega_0, I_0)$ de los elementos orbitales, darán lugar a las correspondientes efemérides $\theta_c = (\theta_k)_0 = f[(q_i)_0, t_k]$.

Desarrollando la función $f(q_i, t_k)$ en el entorno de valores $\{(q_i)_0, t_k\}$, por medio de una serie de Taylor, en el que despreciamos las derivadas parciales de orden superior al primero, tendremos:

$$\sum_{i=1}^{6} (\partial f/\partial q_i)\Delta q_i = \theta_k - \theta_c$$

Y el sistema de estas n ecuaciones con siete incógnitas, tratado por el método de mínimos cuadrados o de mínimos valores absolutos, nos llevará a determinar las correcciones Δq_i de los elementos orbitales previos y por tanto $q_i = (q_i)_0 + \Delta q_i$.

Los nuevos valores P , T , e , Ω , ω , I , obtenidos al cabo de un cierto número de iteraciones, determinan la órbita corregida en cuanto a ángulos de posición y habrá de calcularse el semieje mayor a, puesto que las distancias no han intervenido en el cálculo.

Para ello, con los elementos orbitales anteriores, calcularemos las distancias $(\rho_i)_c$, para un supuesto semieje mayor a = 1, y las iremos comparando con las distancias observadas ρ_i. El promedio aritmético de los cocientes

$$\alpha_i = \frac{\rho_i}{(\rho_i)_c}$$

nos dará el nuevo semieje mayor a en la forma

$$a = \frac{1}{m}\sum_k \alpha_i$$

con lo cual queda terminado el proceso de cálculo.

Entre los posibles métodos de cálculo que determinan la corrección de órbitas, utilizando solamente diferencias $\Delta\theta$, podemos citar los siguientes:

Método de Comstock.

Este método de corrección de órbitas consiste en obtener, para cada observación, los coeficientes A, B, C, D, H, K, que llevados a las ecuaciones

$$A\Delta\Omega + B\Delta I + C\Delta\omega + D\Delta e + Hn\Delta T + K\Delta n = \Delta\theta$$

permiten obtener las correcciones $\Delta\Omega$, ΔI, $\Delta\omega$, Δe, ΔT, Δn, que designaremos en general por $\Delta\sigma$, en función de las diferencias observación-cálculo $\Delta\theta$, obtenidas en una órbita previa de elementos P_0, T_0, a_0, e_0, Ω_0, ω_0, I_0.

Dichos coeficientes vienen determinados por las derivadas parciales $\partial\theta/\partial\sigma_i$, que se obtienen derivando la fórmula

$$\operatorname{tag}(\theta - \Omega) = \operatorname{tag}(\omega + f)\cos I$$

con respecto a cada uno de los seis elementos orbitales. El resultado obtenido, que ha de ser aplicado a cada observación, es el siguiente

$$A = \frac{\partial\theta}{\partial\Omega} = 1$$

$$B = \frac{\partial\theta}{\partial I} = -\frac{1}{2}\operatorname{sen}2(\theta_c - \Omega_0)\ \operatorname{tag} I_0$$

$$C = \frac{\partial\theta}{\partial\omega} = \frac{\operatorname{sen}2(\theta_c - \Omega_0)}{\operatorname{sen}2(\omega_0 + f_c)}$$

$$D = \frac{\partial\theta}{\partial e} = C\left(\frac{1}{1-(e_0)^2} + \frac{a_0}{r_c}\right)\operatorname{sen} f_c$$

$$H = \frac{1}{n}\frac{\partial\theta}{\partial T} = -\frac{(a_0)^2\sqrt{1-(e_0)^2}}{(\rho_c)^2}\cos I_0$$

$$K = \frac{\partial\theta}{\partial n} = -H\,(t_i - T_0)$$

Método de Thiele-Innes modificado.

Veamos como el método de corrección de Thiele-Innes, basado en el cálculo de coordenadas cartesianas, puede ser modificado de manera que se pueda aplicar a la corrección de órbitas utilizando solamente diferencias observación-cálculo en ángulos de posición.

Para ello, recordemos que las coordenadas (x,y), en la órbita aparente, están relacionadas con las cantidades

$$X = X(E, e) = \cos E - e$$

$$Y = Y(E, e) = \sqrt{1-e^2}\ \text{sen}E$$

y las constantes de Innes (A, B, F, G), por medio de las igualdades

$$x = \rho \cos \theta = AX + FY$$

$$y = \rho \text{ sen } \theta = BX + GY$$

de las que se deduce la relación

$$\text{tag } \theta = \frac{BX + GY}{AX + FY}$$

Dicha relación también se puede escribir en la forma

$$\text{tag}\theta = \frac{bX + gY}{X + fY}$$

si se introducen los cocientes

$$b = B/A\ , \qquad f = F/A\ , \qquad g = G/A$$

Partiendo de una órbita previa, de elementos orbitales conocidos, para cada época t_i de observación, pueden ser calculadas las anomalías M_i y E_i, y por tanto tendremos

$$\theta_i = \theta_i[b, f, g, X_i(E, e), Y_i(E, e)]$$

siendo E, a su vez, una función de los elementos orbitales P y T.

Diferenciando dicha función y ordenando el resultado, se obtiene

$$\Delta\theta = \frac{A}{\rho}\cos\theta\,(X\Delta b + Y\Delta g) - \frac{A}{\rho}Y\,\mathrm{sen}\,\theta\,\Delta f + \frac{A^2}{\rho^2}(bf-g)(Y\Delta X - X\Delta Y)$$

Por tanto, si tenemos en cuenta las fórmulas

$$\frac{\partial X}{\partial e} = -1 - \frac{Y^2}{N(1-e^2)} \qquad \frac{\partial Y}{\partial e} = \frac{XY}{N(1-e^2)}$$

$$\frac{\partial X}{\partial M} = -\frac{Y}{N\sqrt{1-e^2}} \qquad \frac{\partial Y}{\partial M} = \frac{X+e}{N}\sqrt{1-e^2}$$

$$N = 1 - e^2 - eX$$

y las igualdades

$$P = -Y\left(1 + \frac{X^2+Y^2}{N(1-e^2)}\right)$$

$$Q = \frac{Y^2 + X(1-e^2)(X+e)}{N\sqrt{1-e^2}}$$

$$R = -Q(t-T)$$

resulta finalmente la relación diferencial

$$\Delta\theta = \frac{A}{\rho}\cos\theta\,(X\Delta b + Y\Delta g) - \frac{A}{\rho}Y\,\mathrm{sen}\,\theta\,\Delta f - \frac{a^2\cos I}{\rho^2}(P\Delta e + Qn\Delta T + R\Delta n)$$

que puede ser tratada por mínimos cuadrados, calculando de esta forma los incrementos de b, f, g, e, T y n.

Por último, para obtener los valores de Ω, ω, I, deberán ser aplicadas las fórmulas (1.4), teniendo en cuenta lo que allí se ha dicho, para poder discriminar los verdaderos valores de los elementos Ω, ω, si bien en este caso han de ser modificadas empleando los coeficientes b, f y g, en lugar de las constantes A, B, F, G, que allí figuran. Es decir que, en su lugar, deberán ser utilizadas las fórmulas siguientes:

$$\mathrm{tag}(\omega + \Omega) = (b - f)/(1 + g)$$

$$\mathrm{tag}(\omega - \Omega) = -(b + f)/(1 - g)$$

$$tag^2\left(\frac{I}{2}\right)=\left(\frac{1-g}{1+g}\right)\frac{\cos(\omega+\Omega)}{\cos(\omega-\Omega)}=-\left(\frac{b+f}{b-f}\right)\frac{\operatorname{sen}(\omega+\Omega)}{\operatorname{sen}(\omega-\Omega)}$$

Método de R. Cid y C. Longás por series de Fourier de la anomalía media.

Veamos un resumen del método de corrección de órbitas elípticas, por medio de series de Fourier, en el que se utiliza la anomalía media M como variable fundamental, que fué publicado en la Rev. de la Academia de Ciencias de Zaragoza (R. Cid y C. Longás, 1992).

En dicho trabajo se demuestra la convergencia uniforme de los desarrollos

$$\frac{C}{n\rho^2}=1+\sum_k(a_k\operatorname{sen}kM+b_k\cos kM)$$

$$\theta=\theta_0+M+\sum_k\left((a_k/k)(1-\cos kM)+(b_k/k)\operatorname{sen}kM\right)$$

así como la de cualquier serie parcial del desarrollo de θ. Diferenciando este desarrollo, tenemos

$$\Delta\theta=\Delta\theta_0+\sum_k\left((\Delta a_k/k)(1-\cos kM)+(\Delta b_k/k)\operatorname{sen}kM\right)-\frac{C}{n\rho^2}\left(\frac{M\Delta P}{P}+n\Delta T\right)$$

En estas condiciones, dado un conjunto m de observaciones θ_α de un par visual y suponiendo calculada una órbita previa de elementos (P_c, T_c, a_c, e_c, Ω_c, ω_c, I_c), podemos calcular las constantes

$$n_c \qquad C_c=n_c a_c^2\sqrt{1-e_c^2}\cos I_c \qquad (A_c, B_c, F_c, G_c)$$

y las efemérides ρ_c, θ_c, para las distintas épocas de observación, así como el valor $(\theta_0)_c$, correspondiente a la anomalía media $M_c = 0$.

En estas condiciones, aplicando, por ejemplo, un proceso de mínimos cuadrados a dicha relación diferencial ecuación, con un cierto número de coeficientes Δa_k, Δb_k, que se extienda al conjunto m de observaciones disponibles, obtendremos, por medio del sistema de ecuaciones normales de Gauss, los incrementos

$$\Delta\theta_0 \quad , \quad \Delta a_k \quad , \quad \Delta b_k \quad , \quad \Delta P \quad , \quad \Delta T \; .$$

cuyos coeficientes respectivos, son:

$$1 \quad , \quad \frac{1}{k}(1 - \cos kM_c) \quad , \quad \frac{1}{k}\operatorname{sen} kM_c \quad , \quad -\frac{C_c M_c}{n_c \rho_c^2 P_c} \quad , \quad -\frac{C_c}{\rho_c^2}$$

siendo $\Delta\theta = \theta - \theta_c$, las diferencias observación-cálculo (O-C), en cada una de las observaciones.

De esta forma se obtendrán los nuevos elementos

$$\theta_0 = (\theta_0)_c + \Delta\theta_0 \quad , \quad P = P_c + \Delta P \quad , \quad T = T_c + \Delta T$$

junto con los incrementos Δa_k , Δb_k , cuyo único interés reside en que sirven de control a los cálculos.

Para obtener la corrección Δe, de la excentricidad, consideremos el valor particular de la anomalía verdadera $f_1 = \pi/2$, para el cual la anomalía excéntrica correspondiente, E_1 será $E_1 = \phi$, siendo $e = \cos\phi$.

De acuerdo con esto, la igualdad

$$\Delta e = \frac{1}{2\sqrt{1-e^2}}\left((\phi - \operatorname{sen}\phi\cos\phi)\frac{\Delta P}{P} + n\Delta T\right)$$

nos permitirá calcular Δe.

Debemos observar que el valor obtenido por esta igualdad viene dado en radianes, por lo cual el resultado ha de ser dividido por 2π para que sea reducido a la unidad.

De esta forma se obtiene un nuevo valor de la excentricidad

$$e = e_c + \frac{1}{2\pi}\Delta e$$

Recordemos ahora que la obtención de los coeficientes Δa_k, Δb_k, solamente es empleado como control de los resultados, por lo cual la resolución del sistema de ecuaciones de Gauss se simplifica notablemente si se introducen unos coeficientes Δa_0, Δb_0, que verifiquen las igualdades

$$\sum_k (1/k)(1-\cos kM)\Delta a_k = \Delta a_0 \sum_k (1/k)(1-\cos kM)$$

$$\sum_k (1/k)\operatorname{sen} kM\Delta b_k = \Delta b_0 \sum_k (1/k)\operatorname{sen} kM$$

ya que en este caso el sistema de ecuaciones de Gauss contiene solamente cinco incógnitas.

En esencia, dicha transformación es equivalente a utilizar unos incrementos Δa_0 , Δb_0 , que son los promedios ponderados de los incrementos Δa_k , Δb_k .

Una vez calculados los nuevos elementos θ , P , T , e , podemos proceder al cálculo de los elementos Ω , ω , I , calculando las anomalías verdaderas f_k , para cada instante t_k de observación y diferenciando con f_k constante la fórmula

$$\operatorname{tag}(\theta_k - \Omega_c) = \operatorname{tag}(\omega_c + f_k)\cos I_c$$

Tendremos así

$$\frac{\Delta\theta_k - \Delta\Omega}{\cos^2(\theta_k - \Omega_c)} = \frac{\Delta\omega \cos I_c}{\cos^2(\omega_c + f_k)} - \Delta I \operatorname{tag}(\omega_c + f_c) \operatorname{sen} I_c$$

o bien

$$\Delta\Omega + R_k\Delta\omega - S_k\Delta I = \Delta\theta_k$$

siendo

$$R_k = \frac{\cos^2(\theta_k - \Omega_c)\cos I_c}{\cos^2(\omega_c + f_k)} \qquad S_k = \operatorname{tag}(\omega_c + f_k)\operatorname{sen} I_c \cos^2(\theta_k - \Omega_c)$$

Tratada esta ecuación por mínimos cuadrados, nos proporcionará los incrementos $\Delta\Omega$, $\Delta\omega$, ΔI, con lo cual

$$\Omega = \Omega_c + \Delta\Omega \qquad \omega = \omega_c + \Delta\omega \qquad I = I_c + \Delta I$$

El proceso es, evidentemente, iterativo y puede ser repetido hasta que sea obtenido un conjunto aceptable de diferencias observación-cálculo en ángulos de posición.

Con los nuevos valores de los elementos P , T , e , Ω , ω , I , obtenidos al cabo de un cierto número de iteraciones, se puede proceder al cálculo del semieje mayor de la forma que vimos al comienzo de este párrafo, quedando, por tanto, terminado el proceso de cálculo.

Bibliografía

Cid, R. (1958): Astronomical Journal, 63, 395. U.S.A.

Cid, R. (1960): Urania nº 252, 129. Tarragona.

Cid, R. (1962): Rev. Acad. de Ciencias de Zaragoza, s. 2, XV, 37.

Cid, R. (1968): Urania nº 267-8, 31. Tarragona.

Cid, R. (1989): Rev. Acad. de Ciencias de Zaragoza, t 44, pags. 109-116.

Cid, R. y Longás, C. (1992): Rev. Acad. de Ciencias de Zaragoza, t. 47, pags. 129-135.

Comstock, G.C. (1918): Astronomical Journal, 31, 33. U.S.A.

Innes, R.T.A. y Van den Bos, W.H. (1926): Union Obs. Cirs., 68, 354 y Union Obs. Cir. 86, 261 (1932).

Longás, C. (1993): Cálculo y corrección de órbitas de estrellas dobles visuales. Tesis. Facultad de Ciencias de Zaragoza.

Thiele, T.N. (1860): Astronomische Nachrichten, 52, 39.

Thiele, T.N. (1883): Astronomische Nachrichten, 104, pag. 245.

Apéndices

Como aplicación de lo expuesto hemos efectuado, por dos de los métodos anteriores, la corrección de la órbita ADS 15972, calculando previamente una órbita provisional por el método de R. Cid. Los resultados obtenidos figuran en los Apéndices I.y II.

APENDICE I

Orbita de la estrellA doble ADS 15972.

Elementos orbitales obtenidos por el método de R. Cid

P = 45.271 años T = 1971.172
a = 2".443 e = 0.405
Ω = 157º.793 ω = 220º.416
I = 160º.182

t	θ	ρ	Observadores	Δθ	Δρ
1930.570	219.50	1.890	VB 4 Kpr 3	.72	.099
31.680	206.20	1.980	VB 4 Sim 3	.01	-.001
32.790	195.00	2.250	Sim 3 GS 3	-.79	.083
33.540	191.40	2.340	VB 4 Kpr 2	1.67	.050
34.740	183.60	2.530	VB 3 Rabe 2 Bz 1	2.37	.054
35.840	176.30	2.560	Rabe 5 VB 3 Bz 3	1.84	-.071
37.730	165.30	2.880	Bz 6 Rabe 3 Kpr 1	.90	.019
37.776	165.52	3.028	pg	1.34	.162
38.719	159.63	2.948	pg	-.09	-.015
38.890	158.70	3.000	Bz 4 Rabe 4	-.24	.021
39.639	156.77	3.055	pg	1.15	.010
39.800	153.90	3.060	Rabe 4 Bz 3	-1.o3	.001
40.660	149.90	3.210	Rabe 5 Bz 4	-1.42	.085
40.776	150.47	3.108	pg	-.37	-.025
41.705	147.83	3.258	pg	.72	.066
41.820	146.70	3.170	Rabe 6 Dur 3 Bz 3 VB 2	.04	-.029
42.776	143.65	3.288	pg	.69	.040
42.870	142.10	3.240	Rabe 7 Bz 3 Dur 2	-.51	-.012
43.764	139.25	3.286	pg	.00	-.001
43.870	139.00	3.250	Rabe 4 Bz 4 VB 4	.15	-.041
44.714	135.75	3.353	pg	.00	.038
45.661	132.48	3.282	pg	.18	-.050
45.800	133.30	3.430	Bz 4 VB 3	1.50	.096
46.760	125.90	3.250	Rabe 5	-2.44	-.091
46.831	128.64	3.451	pg	.56	.110
48.720	121.60	3.350	Rabe 6 VB 4 Bz 3	.34	.025
48.816	120.72	3.343	pg	-.19	.020
49.780	115.20	3.270	Rabe 8 Mark 2	-2.18	-.030
50.780	110.10	3.030	Rabe 5	-3.56	-.238
51.662	109.98	3.235	pg	-.33	.004
51.810	107.20	3.060	Rabe 6 Bz 5 Mark 3	-2.54	-.165
52.684	105.99	3.155	pg	-.32	-.026
53.510	99.50	3.150	Rabe 16 Dju 3	-3.49	.017
53.737	101.66	3.112	pg	-.39	-.007
54.768	97.77	3.140	pg	.07	.090
55.450	92.10	3.000	Rabe 16 Wor 5	-2.61	.000
55.761	93.07	2.960	pg	-.24	-.016
56.721	88.78	2.897	pg	-.06	.000
56.820	87.70	2.820	Wor 5 C 3	-.66	-.069
57.690	83.50	2.740	Wor 4 C 3 B 3	-.56	-.072
58.729	79.02	2.669	pg	.43	-.044
58.810	78.70	2.610	C 3 Wor 1	.55	-.095
59.688	73.14	2.583	pg	-.02	-.033
59.780	72.20	2.590	Wor 8 C 4 hz 4	-.42	-.017
60.777	67.17	2.452	pg	.69	-.049
60.920	66.80	2.420	Wor 5 C 3	1.25	-.065
61.740	58.90	2.330	Wor 4 B 4	-1.12	-.064
61.772	59.42	2.299	pg	-.37	-.092

62.660	52.80	2.150	B 8	Wor 4	C 3	-.48	-.140
62.830	50.90	2.146	pg			-1.07	-.124
63.570	44.70	2.060	Wor 4			-1.27	-.125
63.700	42.03	2.012	pg			-2.84	-.157
64.700	35.80	2.100	Wor 4			-.05	.047
64.753	31.13	1.870	pg			-4.21	-.176
65.800	20.90	1.880	Wor 3			-3.76	-.044
65.861	18.42	1.741	pg			-5.57	-.176
66.790	8.00	1.660	Wor 5			-5.23	-.150
67.722	350.32	1.530	pg			-10.78	-.176
67.840	351.30	1.600	Wor 4			-8.16	-.093
68.740	335,50	1.500	Wor 4			-10.58	-.097
69.790	314.10	1.520	Wor 6			-14.32	.022
70.730	295.50	1.520	Wor 6			-15.25	.090
71.660	276.20	1.470	Wor 4			-15.75	.073
72.750	256.20	1.510	Wor 8			-13.44	.093
73.740	238.80	1.660	Wor 4	hz 4		-11.93	.166
74.790	222.60	1.790	Wor 5			-10.70	.163
75.650	212.70	1.920	Wor 4			-8.51	.160
75.830	211.63	1.964	pg			-7.29	.175
76.786	202.44	2.188	pg			-5.46	.237
76.860	201.70	2.120	Wor 4			-5.43	.157
77.699	195.17	2.301	pg			-3.81	.195
77.700	183.70	2.350	Wor 4			-5.27	.243
78.750	187.30	2.390	Wor 4	hz 3		-2.89	.109
78.759	186.99	2.469	pg			-3.13	.187
79.830	179.30	2.630	Wor 4			-3.13	.181
79.833	179.77	2.544	pg			-2.64	.094
80.670	174.50	2.700	Wor 4	hz 2		-2.58	.129
80.817	173.21	2.714	pg			-2.98	.123
81.738	168.58	2.809	pg			-2.35	.096
82.600	166.80	3.000	hz 2			.40	.184
84.620	154.50	3.099	hz 2			-2.39	.070
84.825	155.46	3.137	pg			-.53	.099
85.509	151.55	3.099	pg			-1.52	.005

APENDICE II

Corrección de la órbita de la estrella doble ADS 15972

α) **Corrección por el método de Thiele modificado con cuatro iteraciones.**

β) **Corrección por series de Fourier de la anomalía media, con coeficientes a_1, b_1, a_2, b_2, y cuatro iteraciones.**

Elementos orbitales corregidos

Con el método de Thiele	Por series de Fourier
P = 44ª.689	P = 44ª.678
T = 1970.304	T = 1970.399
a = 2".3886	a = 2".4303
e = 0.410	e = 0.395
Ω = 155º.595	Ω = 142º.828
ω = 212º.646	ω = 200º.681
I = 167º.994	I = 162º.339

t	Δθ	Δρ	Δθ	Δρ
1930.570	.91	.019	-1.58	.051
1931.680	-.48	-.063	-2.34	-.026
1932.790	-1.62	.036	-2.95	.073
1933.540	.75	.015	-.28	.049
1934.740	1.43	.036	.76	.064
1935.840	.99	-.076	.56	-.056
1937.730	.28	.029	.12	.036
1937.776	.73	.173	.58	.179
1938.719	-.57	.001	-.64	.001
1938.890	-.69	.037	-.75	.036
1939.639	.80	.027	.79	.021
1939.800	-1.35	.019	-1.36	.012
1940.660	-1.62	.104	-1.59	.092
1940.776	-.56	-.006	-.52	-.018
1941.705	.65	.083	.72	.067
1941.820	-.01	-.011	.06	-.028
1942.776	.75	.055	.83	.035
1942.870	-.43	.003	-.35	-.018
1943.764	.17	.010	.27	-.013
1943.870	.33	-.030	.42	-.053
1944.714	.27	.046	.37	.021
1945.661	.52	-.048	.62	-.073
1945.800	1.86	.098	1.96	.072
1946.760	-2.01	-.094	-1.91	-.120
1946.831	.99	.106	1.09	.081
1948.720	.87	.010	.98	-.013
1948.816	.34	.004	.45	-.019
1949.780	-1.62	-.052	-1.49	-.073
1950.780	-2.97	-.265	-2.83	-.282
1951.662	.27	-.028	.43	-.042
1951.810	-1.94	-.197	-1.77	-.210

1952.684	.27	-.061	.47	-.071
1953.510	-2.90	-.021	-2.66	-.027
1953.737	.19	-.046	.44	-.050
1954.768	.63	.050	.94	.051
1955.450	-2.06	-.040	-1.70	-.035
1955.761	.30	-.055	.69	-.049
1956.721	.46	-.037	.94	-.027
1956.820	-.15	-.106	.35	-.094
1957.690	-.07	-.104	.54	-.089
1958.729	.92	-.069	1.69	-.049
1958.810	1.05	-.119	1.83	-.099
1959.688	.50	-.048	1.45	-.025
1959.780	.11	-.030	1.07	-.007
1960.777	1.31	-.048	2.50	-.024
1960.920	1.89	-.063	3.11	-.039
1961.740	-.32	-.048	1.11	-.024
1961.772	.43	-.075	1.87	-.051
1962.660	.62	-.105	2.29	-.084
1962.830	.10	-.086	1.83	-.066
1963.570	.28	-.069	2.22	-.054
1963.700	-1.20	-.099	.77	-.085
1964.700	2.43	.130	4.64	.134
1964.753	-1.68	-.092	.55	-.090
1965.800	.14	.064	2.50	.052
1965.861	-1.57	-.067	.78	-.080
1966.790	.53	-.024	2.78	-.054
1967.722	-2.70	-.043	-.89	-.090
1967.840	.25	.040	1.97	-.009
1968.740	.53	.025	1.38	-.036
1969.790	-.09	.107	-.70	.041
1970.730	.89	.119	-1.08	.061
1971.660	.66	.035	-2.30	-.005
1972.750	1.17	-.017	-2.21	-.030
1973.740	-.05	.012	-3.21	.021
1974.790	-1.71	-.009	-4.33	.016
1975.650	-1.45	-.011	-3.58	.021
1975.830	-.57	.005	-2.60	.038
1976.786	-.27	.080	-1.80	.116
1976.860	-.33	.001	-1.83	.037
1977.699	.36	.054	-.78	.089
1977.700	-1.10	.103	-2.24	.137
1978.750	.48	-.010	-.30	.020
1978.759	.23	.068	-.54	.098
1979.830	-.28	.083	-.78	.107
1979.833	.21	-.003	-.29	.020
1980.670	.01	.047	-.32	.064
1980.817	-.42	.043	-.73	.060
1981.738	.03	.031	-.14	.041
1982.600	2.69	.131	2.60	.135
1984.620	-.18	.039	-.12	.030
1984.825	1.68	.070	1.75	.060
1985.509	.71	-.018	.79	-.031

Rev. Academia de Ciencias. Zaragoza. 48 (1993)

COMPARACIÓN DE ALGORITMOS DE TRANSFORMACIÓN DE COORDENADAS GEOCÉNTRICAS A GEODÉSICAS

Roberto Barrio y Andrés Riaguas

Grupo de Mecánica Espacial, Departamento de Matemática Aplicada. Universidad de Zaragoza

Resumen: En esta nota analizamos distintos métodos para realizar el cambio entre coordenadas geocéntricas y geodésicas con vistas a su aplicación en geodesia espacial. Se comparan fórmulas aproximadas y métodos iterativos considerando dos altitudes (956 y 21685 kilómetros). Concluímos recomendando el método iterativo de Bowring debido a su alta precisión y rapidez.

1. Introducción

La transformación entre coordenadas geodésicas y geocéntricas es una de las tareas más frecuentes en los cálculos geodésicos. La necesidad de seleccionar un algoritmo para su implementación en un software de geodesia espacial nos llevó a una comparación numérica de su precisión y rapidez entre varios de ellos.

Las expresiones que relacionan las coordenadas geocéntricas y las geodésicas son las siguientes:

$$\rho \cos \phi' = (C + h) \cos \phi, \tag{1}$$

$$\rho \operatorname{sen} \phi' = (S + h) \operatorname{sen} \phi, \tag{2}$$

donde:

$$C = C(e, \phi) = a(1 - e^2 \operatorname{sen}^2 \phi)^{-\frac{1}{2}}, \qquad S = S(e, \phi) = C(1 - e^2). \tag{3}$$

Mientras que para obtener las coordenadas geocéntricas a partir de las geodésicas es un cálculo directo, la transformación inversa requiere el empleo de otras técnicas como procesos iterativos o fórmulas aproximadas procedentes de series de potencias. Los métodos que a continuación se comparan son los de Morrison y Pines, Long, Deprit y Deprit-Bartholome, GEODYN y Bowring.

Recogemos aquí la notación empleada por nosotros para todos los métodos: ρ distancia geocéntrica medida en radios terrestres, ϕ' latitud geocéntrica, (X, Y, Z) coordenadas cartesianas, h altitud geodésica sobre la superficie del elipsoide, ϕ latitud geodésica, a semieje mayor de la elipse terrestre meridional, e excentricidad del elipsoide, e' segunda excentricidad del elipsoide, b radio polar y f aplanamiento terrestre. A lo largo del artículo se ha tomado como unidad de longitud el radio ecuatorial medio, es decir, $a = 1$ salvo cuando a aparezca explícitamente.

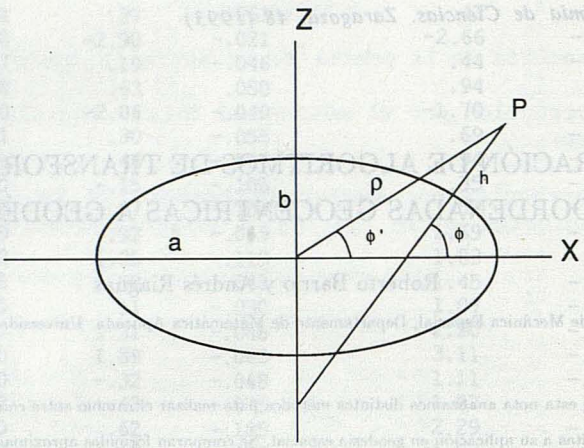


Figura 1.—Diagrama de latitudes geocéntricas y geodésicas donde se ha realizado una sección por un meridiano al elipsoide de revolución.

2. Métodos

2.1 Método de Morrison y Pines

Este método se basa en la fórmula de Lagrange para obtener desarrollos en serie. La solución viene dada como una serie de Fourier de ϕ hasta el orden de e^8.

Partimos de las ecuaciones que relacionan las coordenadas geodésicas con las geocéntricas (1), (2). Podemos expresar estas ecuaciones de la forma:

$$\mathrm{tg}\phi = \mathrm{tg}\phi' + \frac{e^2}{\rho\cos\phi'}\frac{\phi}{(1-e^2\,\mathrm{sen}^2\,\phi)^{\frac{1}{2}}}.$$

Usando la fórmula de Lagrange (ver Whittaker y Watson[9], pág. 132-133) obtendremos:

$$\boxed{\phi = \phi' + a_2\,\mathrm{sen}\,2\phi' + a_4\,\mathrm{sen}\,4\phi' + a_6\,\mathrm{sen}\,6\phi' + a_8\,\mathrm{sen}\,8\phi'} \tag{4}$$

Donde los coeficientes $a_i = a_i(e,\rho)$ de la ecuación anterior son:

$$\begin{aligned} a_2 &= \frac{1}{1024\rho}(512e^2 + 128e^4 + 60e^6 + 35e^8) + \frac{1}{32\rho^2}(e^6 + e^8) \\ &\quad -\frac{3}{256\rho^3}(4e^6 + 3e^8) \\ a_4 &= \frac{-1}{1024\rho}(64e^4 + 48e^6 + 35e^8) + \frac{1}{16\rho^2}(4e^4 + 2e^6 + e^8) \\ &\quad +\frac{15e^8}{256\rho^3} - \frac{e^8}{16\rho^4} \end{aligned} \tag{5}$$

$$\boxed{\begin{aligned} a_6 &= \frac{3}{1024\rho}(4e^6+5e^8)-\frac{3}{32\rho^2}(e^6+e^8) \\ &\quad +\frac{35}{768\rho^3}(4e^6+3e^8) \\ a_8 &= \frac{e^8}{2048}\left(-\frac{5}{\rho}+\frac{64}{\rho^2}-\frac{252}{\rho^3}+\frac{320}{\rho^4}\right) \end{aligned}} \tag{6}$$

Una vez calculada la latitud geodésica, podemos calcular la altitud h con la siguiente expresión:

$$\boxed{h_{MP} = \rho\cos(\phi-\phi') - (1-e^2\,\mathrm{sen}^2\,\phi)^{\frac{1}{2}}} \tag{7}$$

Morrison y Pines dan otras dos fórmulas para el cálculo de h que hemos analizado, dando menor precisión y singularidades para algunos valores de la latitud geodésica. Por lo tanto sólo consideraremos la fórmula indicada en (7).

2.2 Método de Long

Long obtuvo un conjunto de fórmulas expresadas en desarrollos en serie de potencias del aplanamiento terrestre f y de la excentricidad e. Partiendo de las relaciones (3), $(1-f)^2 = 1-e^2$ y del desarrollo de C hasta f^2:

$$C = 1+\frac{1}{2}f+\frac{5}{16}f^2-\frac{1}{2}(f+f^2)\cos 2\phi+\frac{3}{16}f^2\cos 4\phi+\ldots$$

Long obtiene sus fórmulas usando además relaciones entre triángulos, el hecho de que $\phi-\phi'$ sea del mismo orden de magnitud que f y desarrollando las expresiones en series de potencias de f (reteniendo hasta el segundo orden). Teniendo en cuenta el desarrollo de f en potencias de e truncado en e^4

$$f = \frac{1}{2}e^2+\frac{1}{8}e^4\,,$$

resulta:

$$\boxed{\begin{aligned} \phi &= \phi'+\frac{\mathrm{sen}\,2\phi'}{2\rho}e^2+\left[\frac{\mathrm{sen}\,2\phi'}{8\rho}+\left(\frac{1}{4\rho^2}-\frac{1}{16\rho}\right)\mathrm{sen}\,4\phi'\right]e^4 \\ h_L &= (\rho-1)+\frac{1}{4}(1-\cos 2\phi')e^2+ \\ &\quad +\left[\frac{1}{16}(1-\cos 2\phi')+\left(\frac{1}{16\rho}-\frac{1}{64}\right)(1-\cos 4\phi')\right]e^4 \end{aligned}} \tag{8}$$

2.3 Método de Deprit y Deprit-Bartholome

En [3] se describe la forma de obtener por métodos recurrentes esta transformación con la precisión que se desee; en la práctica para evitar todos estos cálculos en cada punto se indican los coeficientes para un determinado valor de f del desarrollo en serie de ϕ y h. Se necesitan

también las series de las derivadas parciales de h y ϕ con respecto de f para que dichas fórmulas se puedan aplicar a otros valores de f posteriores. El cálculo de ϕ y de h quedaría:

$$\phi + \frac{\partial \phi}{\partial f}\Delta f, \qquad h + \frac{\partial h}{\partial f}\Delta f. \tag{9}$$

A continuación se muestran en una tabla las siguientes series ($u = 1/\rho$):

$$h = \sum_{n\geq 0} h_n(\phi')u^n, \qquad \frac{\partial h}{\partial f} = \sum_{n\geq 0} h_n^*(\phi')u^n, \tag{10}$$

$$\phi = \sum_{n\geq 0} \phi_n(\phi')u^n, \qquad \frac{\partial \phi}{\partial f} = \sum_{n\geq 0} \phi_n^*(\phi')u^n. \tag{11}$$

Con los coeficientes para el valor de f igual a 1/298.25 (adoptado por la IAU en 1968) y con una precisión de 9 dígitos significativos, Deprit y Deprit-Bartholome dan la siguiente tabla

	h (en cosenos)	$\partial h/\partial f$	ϕ (en senos)	$\partial\phi/\partial f$
$0\phi'$	$+1.000000000\rho$			
	-0.998324258	-0.000005609		
	$+0.000002810\rho^{-1}$	$-0.000000009\rho^{-1}$		
$2\phi'$	-0.001676445	$+0.000005612$		
	$-0.000000005\rho^{-1}$		$+0.003352891\rho^{-1}$	$-0.000011223\rho^{-1}$
	$+0.000000009\rho^{-2}$		$+0.000000009\rho^{-2}$	
			$-0.000000014\rho^{-3}$	
$4\phi'$	$+0.000000704$	-0.000000002		
	$-0.000002810\rho^{-1}$	$+0.000000009\rho^{-1}$	$-0.000002815\rho^{-1}$	$+0.000000009\rho^{-1}$
			$+0.000011242\rho^{-2}$	$-0.000000038\rho^{-2}$
$6\phi'$	$+0.000000005\rho^{-1}$		$+0.000000004\rho^{-1}$	
	$-0.000000009\rho^{-2}$		$-0.000000028\rho^{-2}$	
			$+0.000000055\rho^{-3}$	

2.4 GEODYN II

Este método se basa en un proceso iterativo dado por Heiskanen y Moritz[8]. Para ello se supondrá que $h \ll C$, no obstante numéricamente es aceptable para altitudes mayores. Sean X, Y, Z las coordenadas cartesianas de un punto P respecto de un sistema con el origen en el centro de la tierra y eje polar Z, t y Z_t las indicadas en la figura (2). Podemos aproximar:

$$C \operatorname{sen}\phi \simeq Z.$$

De la figura (2) se tiene que:

$$t = Ce^2 \operatorname{sen}\phi.$$

Podemos elegir entonces como valor inicial para la iteración:

$$t_0 = e^2 Z.$$

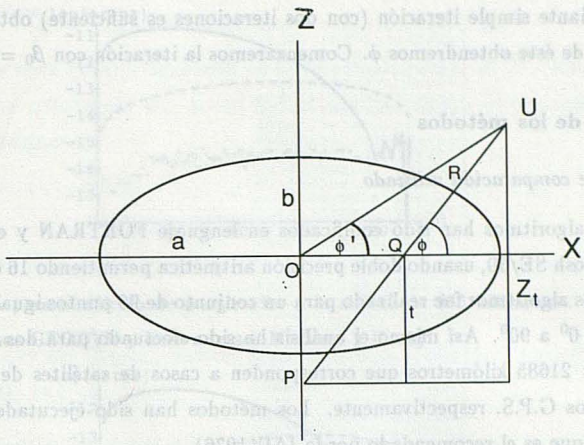


Figura 2.—Diagrama de latitudes geocéntricas y geodésicas donde se ha realizado una sección por un meridiano al elipsoide. $\overline{PQ} = Ce^2$, $\overline{PR} = C$, $\overline{RU} = h$, $\overline{QR} = S$, $\overline{OU} = \rho$.

Se plantea el siguiente esquema para la iteración:

$$\begin{aligned} Z_t &= Z + t \\ C + h &= (X^2 + Y^2 + Z_t^2)^{\frac{1}{2}} \\ \operatorname{sen}\phi &= \frac{Z_t}{C+h} \\ C &= \frac{1}{\sqrt{1 - e^2 \operatorname{sen}^2 \phi}} \\ t &= Ce^2 \operatorname{sen}\phi \end{aligned} \tag{12}$$

2.5 Método de Bowring

Bowring indicó un algoritmo basado en el método de la simple iteración. Haciéndonos eco del artículo de Laskowski comentamos este método y no el más reciente de Borkowski(1989) (basado en el método de Newton-Raphson). Laskowski nos muestra numéricamente que en este caso funciona mejor la simple iteración. Por lo tanto, al obtener resultados más rápidos y precisos, sólo comentaremos este método.

De la ecuación de la latitud expandida en series de Taylor y puesta en una forma adecuada nos resulta:

$$\operatorname{tg}\beta = \frac{b(Z + be'^2 \operatorname{sen}^3\beta)}{a(p - ae^2 \cos^3\beta)} \tag{13}$$

donde:

$$e^2 = (a^2 - b^2)/a^2, \qquad e'^2 = (a^2 - b^2)/b^2,$$

$$\operatorname{tg}\beta = \frac{b}{a}\operatorname{tg}\phi, \qquad p = (X^2 + Y^2)^{\frac{1}{2}}.$$

Ahora mediante simple iteración (con dos iteraciones es suficiente) obtendremos una estimación de β y de éste obtendremos ϕ. Comenzaremos la iteración con $\beta_0 = \text{arctg}\left[\frac{aZ}{bp}\right]$.

3. Análisis de los métodos

3.1 Método de comparación utilizado

Todos los algoritmos han sido codificados en lenguaje FORTRAN y ejecutados en ordenadores Macintosh SE/30, usando doble precisión aritmética permitiendo 16 cifras significativas. El análisis de los algoritmos fue realizado para un conjunto de 90 puntos igualmente distribuidos en latitudes de 0^0 a 90^0. Así mismo el análisis ha sido efectuado para dos altitudes distintas: $h = 956$ y $h = 21685$ kilómetros que corresponden a casos de satélites de altitudes cercanas al Starlette y los G.P.S. respectivamente. Los métodos han sido ejecutados para el valor de $f = 1/298.257$, que es el recomendado por la IAU(1976).

El error de la transformación es definido como la distancia euclídea entre la posición geodésica verdadera (la introducida por nosotros) y la calculada a partir de la posición geocéntrica. Dicha posición se obtiene del paso de la posición geodésica verdadera (el error de esta conversión se supone nulo). En las gráficas se ha representado el logaritmo decimal del valor absoluto del error contra el ángulo en grados, con lo que el eje de abcisas representa la latitud. Para los casos en que por falta de precisión del ordenador obtengamos error cero lo representaremos en las gráficas por 10^{-18}. Del método de Bowring, al presentar un mejor comportamiento, se ha realizado un diagrama más completo.

3.2 Análisis

- Los métodos de Long y Deprit y Deprit-Bartholome fueron construidos para satisfacer las exigencias en precisión que se tenía cuando fueron publicados. Así para el método de Long hemos obtenido errores de hasta 10^{-7} (debido a que en el desarrollo se trunca en e^4) y para el segundo sólo se ha tenido en cuenta el desarrollo ya calculado para la f indicada que fue construido para obtener un error del orden de 10^{-10} (se ha comprobado numéricamente dicha precisión). Por estas razones nos centramos en los restantes métodos.
- Las siguientes gráficas (figuras (3), (4), (5) y (6)) muestran las diferencias de los errores de los restantes métodos. En las dos gráficas ((3), (4)) se analizan los errores en la obtención de ϕ para las dos altitudes mencionadas. Hemos de notar que en los dos procesos iterativos se aprecian dos zonas en las cuales se incrementa la precisión, llegando en algunos casos al umbral de precisión de la máquina, cuando se calcula ϕ. Estas zonas corresponden a valores de ϕ cercanos a 5° y a 80°. En las figuras ((5), (6)) se comparan sólo dos métodos de obtener h, GEODYN y Bowring. Este último se realiza usando h_{MP} de la fórmula (5) y por lo tanto la gráfica para el método de Morrison y Pines es análoga. Los métodos iterativos, del GEODYN y Bowring, han sido ejecutados con 4 y 2 iteraciones respectivamente.

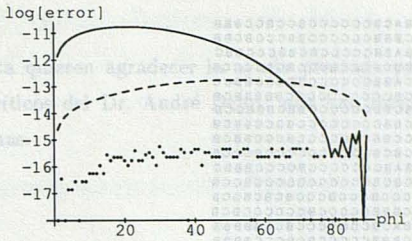


Figura 3.—Métodos para el cálculo de ϕ aplicados a la altitud de 956 kilómetros. (línea punteada Bowring, línea continua GEODYN y línea a trazos Morrison y Pines)

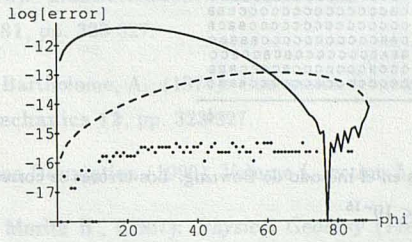


Figura 4.—Métodos para el cálculo de ϕ aplicados a la altitud de 21685 kilómetros. (línea punteada Bowring, línea continua GEODYN y línea a trazos Morrison y Pines)

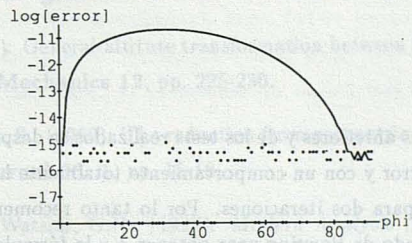


Figura 5.—Métodos para el cálculo de h aplicados a la altitud de 956 kilómetros. (línea punteada h_{MP} de la equación (5) partiendo de ϕ obtenido de Bowring y línea continua GEODYN)

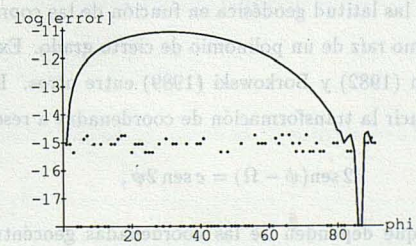


Figura 6.—Métodos para el cálculo de h aplicados a la altitud de 21685 kilómetros. (línea punteada h_{MP} de la equación (5) partiendo de ϕ obtenido de Bowring y línea continua GEODYN)

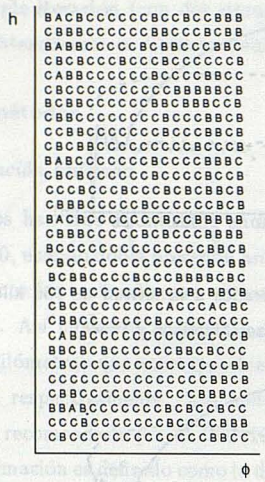


Figura 7.—Diagrama de errores en el método de Bowring. Los errores se corresponden con: $A < 10^{-17}$, $10^{-17} < B < 10^{-16}$, $10^{-16} < C < 10^{-15}$

- Realizamos un análisis más exhaustivo del método de Bowring debido a su mayor precisión. El diagrama (7) nos muestra la semisuma de los errores para ϕ y h. El valor de h varía de 32 en 32km., con un valor inicial de 32km. y ϕ varía de 5 en 5 grados de 0° a 90°.

4. Conclusión

Del análisis de las gráficas anteriores y de los tests realizados se desprende que el método de Bowring presenta el menor error y con un comportamiento totalmente homogéneo. Dicho error se sitúa alrededor de 10^{-16} para dos iteraciones. Por lo tanto recomendamos de entre todos estos métodos el uso del método de Bowring para obtener ϕ y la fórmula (7) para h.

5. Apéndice

Existe otra forma de obtener las latitud geodésica en función de las coordenadas geocéntricas y es obteniendo las primera como raíz de un polinomio de cierto grado. Expresiones cerradas han sido obtenidas por Heikkinen (1982) y Borkowski (1989) entre otros. La forma de obtener la fórmula de Borkowski es reducir la transformación de coordenadas a resolver la ecuación en ψ:

$$2\,\mathrm{sen}(\psi - \Omega) = c\,\mathrm{sen}\,2\psi\,,$$

siendo c, Ω y ψ cantidades que dependen de las coordenadas geocéntricas, geodésicas y del elipsoide de referencia. Mediante el cambio de variable $t = \mathrm{tg}(\frac{\pi}{4} - \frac{\psi}{2})$ se obtiene una ecuación de cuarto grado cuyas soluciones se pueden expresar de forma cerrada. Los inconvenientes de estas fórmulas son tanto las operaciones complejas como la elección de la solución correcta que llevan a la pérdida de cifras significativas y elevan el tiempo de cálculo.

6. Agradecimientos

Los autores de esta nota quieren agradecer la ayuda prestada por el Dr. Sebastián Ferrer y los consejos y juicios críticos del Dr. André Deprit que han mejorado de forma ostensible el contenido de estas páginas.

Referencias

[1] Borkowski, K.M., (1989): Accurate algorithms to transform geocentric to geodetic coordinates, **Bulletin Geodesique 63**, pp. 50-56.

[2] Bowring, B.R., (1976): Transformation from spatial to Geografical Coordinates, **Survey Review, XXIII, 181**, pp. 323-327.

[3] Deprit, A. y Deprit-Bartholome, A., (1974): Conversion from geocentric to geodetic coordinates, **Celestial Mechanics 12**, pp. 323-327.

[4] GEODYN II, Systems Description (1990), Volume I, section 5.0

[5] Heiskanen W. A. y Moritz H., (1967): *Physical Geodesy* (Freeman , W. H. and Co., San Francisco).

[6] Laskowski, P., (1991): Is Newton's iteration faster than simple iteration for transformation between geocentric and geodetic coordinates?, **Bulletin Geodesique 65**, pp. 14-17.

[7] Long, S.A.T., (1974): General-altitute transformation between geocentric and geodetic coordinates, **Celestial Mechanics 12**, pp. 225-230.

[8] Morrison, J. y Pines, S., (1960): The reduction from geocentric to geodetic coordinates, **The Astronomical Journal 66, 1**, pp. 15-16.

[9] Whittaker, E.T. y Watson, G.N., (1920): *Modern Analysis* (Cambridge University Press, New York), 3rd ed.

Rev. Academia de Ciencias. Zaragoza. 48 (1993)

UN ESQUEMA MUESTRAL SIN REEMPLAZAMIENTO INMEDIATO

por

M. Ruiz Espejo* y M. Rueda García**

*Departamento de Estadística
Facultad de Ciencias Económicas y Empresariales
Universidad Complutense de Madrid

y

**Departamento de Estadística
Facultad de Ciencias Matemáticas
Universidad de Granada

Summary. A sampling scheme is proposed in the context of sampling theory, which is unbiased for the population mean (with sample mean estimator) and intermediate between simple random sampling with and without replacement in variance and expected cost. This scheme is characterized as a Markov chain.

1. Introducción

Un esquema muestral es un procedimiento probabilístico para seleccionar en primer lugar una unidad i_1 de entre las N que constituyen una población finita, $U = \{1,2,\ldots,N\}$. Seguidamente obtener otra unidad i_2 de entre las N, sabiendo que en primer lugar se obtuvo i_1. Y en una etapa genérica n, seleccionar la unidad i_n sabiendo que previamente se ha obtenido la secuencia

$$\underline{s}_{n-1} = (i_1, i_2, \ldots, i_{n-1}) .$$

Es decir, debe especificarse en general las probabilidades

$$p(X_1 = i_1)$$

y

$$p(X_n = i_n \mid X_1 = i_1, X_2 = i_2, \ldots, X_{n-1} = i_{n-1}) = p(X_n = i_n \mid \underline{s}_{n-1}),\ n = 2,3,4,\ldots$$

donde X_n es la unidad de la población finita U, seleccionada en la etapa n.

Como se justifica en Cassel et al. (1977) es equivalente usar un "esquema muestral" a un "diseño muestral ordenado" que reproduzca sus probabilidades de selección.

Podemos considerar un esquema muestral como una cadena de Markov homogénea (Vélez, 1977). En este caso el espacio de estados será la propia población finita U, de tamaño N. La distribución inicial equiprobable asignaría

$$p(X_1 = i) = 1/N \text{ para } i=1,2,\ldots,N.$$

Y la matriz de probabilidades de transición en una etapa para un esquema muestral "sin reemplazamiento inmediato", o también "con reemplazamiento no inmediato" puede venir expresada por

$$p(X_{n+1}= j \mid X_n= i) = (1 - \delta_{ij})/(N-1) \text{ para } i,j=1,2,\ldots,N,$$

donde δ_{ij} son las deltas de Krönecker (toma el valor 1 si i=j, y el valor 0 si i≠j). De este modo quedaría caracterizado el esquema muestral de Markov, que coincide con el muestreo aleatorio simple sin o con reemplazamiento si el tamaño muestral es n=1, coincide con el muestreo aleatorio simple sin reemplazamiento si n=2, y si n⩾3 sería una estrategia (con la media muestral $\bar{y}_s$) intermedia en cuanto a su varianza y coste esperado que las estrategias de muestreo aleatorio simple con y sin reemplazamiento (Ruiz y Santos, 1989).

Además, al ser un diseño con reemplazamiento (aunque no inmediato), la estrategia propuesta admite otras insesgadas y de menor o igual varianza tal y como es bien sabido (Cassel et al., 1977 ó Ruiz, 1988) pero al precio de un mayor o igual coste esperado.

Considerando como diseño muestral ordenado de tamaño fijo n, si $i_j \neq i_{j+1}$ para j=1,2,...,n-1, la probabilidad de una muestra ordenada $\underline{s}=(i_1,i_2,\ldots,i_n)$ es

$$p(\underline{s}) = \frac{1}{N}\left(\frac{1}{N-1}\right)^{n-1}.$$

Llamando "p" a éste diseño ordenado propuesto, veremos que con la media muestral $\bar{y}_s$ es una estrategia insesgada para la media poblacional

$$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y}_U ,$$

siendo y_i la variable de interés en la unidad i de U. Además su varianza es calculable para un tamaño muestral n, según proponemos, haciendo uso de resultados de procesos estocásticos y concretamente de cadenas de Markov homogéneas.

2. Insesgación

La estrategia propuesta $(p,\bar{y}_s)$ para un tamaño muestral n, es insesgada para estimar la media poblacional. En efecto,

$$E_p(y_{i_1}) = \sum_{i=1}^{N} y_i p(X_1=i) = \sum_{i=1}^{N} y_i \frac{1}{N} = \bar{y}_U = \mu . \qquad (1)$$

Además,

$$E(y_{i_2}|y_{i_1}) = \sum_{i\neq i_1}^{N} y_i p(X_2=i|X_1=i_1) = \sum_{i\neq i_1}^{N} y_i \frac{1}{N-1} = \bar{y}_{U-\{i_1\}} ,$$

de donde

$$E_p(y_{i_2}) = E[E(y_{i_2}|y_{i_1})] = E(\bar{y}_{U-\{i_1\}}) = \sum_{i_1=1}^{N} \bar{y}_{U-\{i_1\}} p(X_1=i_1) =$$

$$= \sum_{i_1=1}^{N} \left(\frac{1}{N-1} \sum_{j\neq i_1}^{N} y_j\right) \frac{1}{N} = \frac{1}{N(N-1)} \sum_{i_1=1}^{N} \sum_{j\neq i_1}^{N} y_j =$$

$$= \frac{1}{N(N-1)} \sum_{i_1=1}^{N} (N\bar{y}_U - y_{i_1}) = \frac{1}{N(N-1)} (N^2\bar{y}_U - N\bar{y}_U) = \bar{y}_U \qquad (2)$$

Análogamente se procede con

$$E(y_{i_j}) = \bar{y}_U = \mu \quad \text{para } j > 2.$$

Por todo ello, de las ecuaciones (1), (2) y ésta última, deducimos que

$$E(p,\bar{y}_s) = E\left(\frac{1}{n} \sum_{j=1}^{n} y_{i_j}\right) = \frac{1}{n} \sum_{j=1}^{n} E(y_{i_j}) = \frac{1}{n} \sum_{j=1}^{n} \mu = \mu .$$

3. Varianza de la estrategia

Podemos expresar

$$V(p,\bar{y}_s) = E_p[(\bar{y}_s - \mu)^2] = E_p(\bar{y}_s^2) - \mu^2 ,$$

por lo que sólo nos queda calcular

$$E_p(\bar{y}_s^2) = E_p\left(\frac{1}{n}\sum_{j=1}^{n} y_{i_j}\right)^2 = \frac{1}{n^2} E_p\left(\sum_{j=1}^{n}\sum_{k=1}^{n} y_{i_j} y_{i_k}\right) =$$

$$= \frac{1}{n^2} E_p\left(\sum_{j=1}^{n} y_{i_j}^2 + 2\sum_{j=1}^{n-1}\sum_{k>j}^{n} y_{i_j} y_{i_k}\right) =$$

$$= \frac{1}{n^2}\left[\sum_{j=1}^{n} E_p(y_{i_j}^2) + 2\sum_{j=1}^{n-1}\sum_{k>j}^{n} E_p(y_{i_j} y_{i_k})\right] \qquad (3)$$

siendo

$$E_p(y_{i_1}^2) = \sum_{i=1}^{N} y_i^2\, p(X_1=i) = \sum_{i=1}^{N} y_i^2 \frac{1}{N} = \alpha_2 , \qquad (4)$$

donde α_2 es el momento poblacional con respecto al orígen de orden 2, y si $j \geq 2$

$$E_p(y_{i_j}^2) = \sum_{i=1}^{N}\sum_{i_j \neq i}^{N} y_{i_j}^2\, p(X_{j-1}=i)\,p(X_j=i_j \mid X_{j-1}=i) =$$

$$= \sum_{i=1}^{N}\sum_{i_j \neq i}^{N} y_{i_j}^2 \frac{1}{N(N-1)} = \frac{1}{N(N-1)} \sum_{i=1}^{N} (N\alpha_2 - y_i^2) =$$

$$= \frac{1}{N(N-1)} (N^2\alpha_2 - N\alpha_2) = \alpha_2 , \qquad (5)$$

de donde de (4) y (5),

$$\sum_{j=1}^{n} E_p(y_{i_j}^2) = n\alpha_2 . \qquad (6)$$

De modo similar,

$$E_p(y_{i_1} y_{i_2}) = E[E(y_{i_1} y_{i_2} | X_1=i_1)] = E[y_{i_1} E(y_{i_2} | X_1=i_1)] =$$

$$= E(y_{i_1} \bar{y}_{U-\{i_1\}}) = \sum_{i=1}^{N} y_i \bar{y}_{U-\{i\}} p(X_1=i) =$$

$$= \sum_{i=1}^{N} y_i \left(\frac{1}{N-1} \sum_{j\neq i}^{N} y_j\right) \frac{1}{N} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j\neq i}^{N} y_i y_j =$$

$$= \frac{1}{N(N-1)} \sum_{i=1}^{N} y_i \left(\sum_{j\neq i}^{N} y_j\right) = \frac{1}{N(N-1)} \sum_{i=1}^{N} y_i (N\bar{y}_U - y_i) =$$

$$= \frac{1}{N(N-1)} (N^2 \bar{y}_U^2 - N\alpha_2) = \frac{N\mu^2 - \alpha_2}{N-1} \qquad (7)$$

y razonando de modo paralelo,

$$E_p(y_{i_j} y_{i_{j+1}}) = E_p(y_{i_1} y_{i_2}) \quad \text{para } j=2,3,\ldots,n-1. \qquad (8)$$

Para calcular

$$E_p(y_{i_1} y_{i_3}) = E[E(y_{i_1} y_{i_3} | X_1=i_1)] = E[y_{i_1} E(y_{i_3} | X_1=i_1)] , \qquad (9)$$

tenemos que si $i_1 \neq i_3$, por la ecuación de Chapman-Kolmogorov

$$p(y_{i_3} | X_1=i_1) = \sum_{i\neq i_1, i_3}^{N} p(X_2=i | X_1=i_1) p(X_3=i_3 | X_2=i) =$$

$$= \sum_{i\neq i_1, i_3}^{N} \frac{1}{N-1} \frac{1}{N-1} = \frac{N-2}{(N-1)^2} , \qquad (10)$$

y si $i_1=i_3$,

$$p(y_{i_3} | X_1=i_1) = \sum_{i\neq i_1}^{N} p(X_2=i | X_1=i_1) p(X_3=i_1 | X_2=i) =$$

$$= \sum_{i\neq i_1}^{N} \frac{1}{N-1} \frac{1}{N-1} = \frac{1}{N-1} \qquad (11)$$

por lo que de (10) y (11) tenemos

$$E(y_{i_3} | X_1=i_1) = \sum_{i\neq i_1}^{N} y_i \frac{N-2}{(N-1)^2} + y_{i_1} \frac{1}{N-1} ,$$

y sustituyéndolo en (9),

$$E_p(y_{i_1} y_{i_3}) = E[y_{i_1} (\sum_{i\neq i_1}^{N} y_i \frac{N-2}{(N-1)^2} + y_{i_1} \frac{1}{N-1})] =$$

$$= \sum_{i_1=1}^{N} y_{i_1} (\sum_{i\neq i_1}^{N} y_i \frac{N-2}{(N-1)^2} + y_{i_1} \frac{1}{N-1}) \frac{1}{N} =$$

$$= \frac{1}{N(N-1)} \sum_{i_1=1}^{N} y_{i_1}^2 + \sum_{i_1=1}^{N} \sum_{i\neq i_1}^{N} y_{i_1} y_i \frac{(N-1)-1}{N(N-1)^2} =$$

$$= \frac{N}{N-1} \frac{1}{N^2} (\sum_{i_1=1}^{N} y_{i_1}^2 + \sum_{i_1=1}^{N} \sum_{i\neq i_1}^{N} y_{i_1} y_i - \frac{1}{N-1} \sum_{i_1=1}^{N} \sum_{i\neq i_1}^{N} y_{i_1} y_i) =$$

$$= \frac{N}{N-1} \frac{1}{N^2} [N^2\mu^2 - \frac{1}{N-1} \sum_{i_1=1}^{N} y_{i_1} (N\mu - y_{i_1})] =$$

$$= \frac{N}{N-1} \mu^2 - \frac{1}{(N-1)^2} (N\mu^2 - \alpha_2) =$$

$$= \frac{N(N-2)}{(N-1)^2} \mu^2 + \frac{1}{(N-1)^2} \alpha_2 . \qquad (12)$$

Además, la fórmula (12) es válida para

$$E_p(y_{i_j} y_{i_{j+2}}) \quad \text{siendo } j=2,3,\ldots,n-2. \qquad (13)$$

De (3), sustituyendo (6), (7) y (8), así como (12) y (13), podemos concluir que

$$E_p(\bar{y}_s^2) = \frac{1}{n^2} \left\{ n\alpha_2 + 2[(n-1) \frac{N\mu^2 - \alpha_2}{N-1} + \right.$$

$$+ (n-2)\frac{N(N-2)\mu^2 + \alpha_2}{(N-1)^2} + \ldots + E_p(y_{i_1} y_{i_n})]\Big\} \qquad (14)$$

donde en general,

$$E_p(y_{i_1} y_{i_n}) = \sum_{i_1=1}^{N} \sum_{i_n=1}^{N} y_{i_1} y_{i_n} p_{i_1 i_n}^{(n-1)} ,$$

siendo $p_{i_1 i_n}^{(n-1)}$ la probabilidad de transición de i_1 a i_n en n-1 etapas, obtenible de la potencia n-1 de la matriz de probabilidades de transición en una etapa. Sustituyendo (14) en

$$V(p,\bar{y}_s) = E_p(\bar{y}_s^2) - \mu^2 ,$$

obtenemos la varianza de la media muestral bajo el esquema muestral sin reemplazamiento inmediato.

Es fácil comprobar que si n=1 obtenemos los diseños de muestreo aleatorio simple con o sin reemplazamiento, si n=2 estamos ante el muestreo aleatorio simple sin reemplazamiento, así como que si $n \geq 3$ tenemos una estrategia intermedia en varianza y coste esperado.

Referencias

|1| Cassel, C.M., Särndal, C.E. y Wretman, J.H. (1977). Foundations of Inference in Survey Sampling, Wiley, Nueva York.

|2| Ruiz, M. (1988). El teorema de Rao-Blackwell en poblaciones finitas. XVII Congreso SEIO (Benidorm, 4 - 8 de abril).

|3| Ruiz, M. y Santos, J. (1989). Estrategias intermedias de muestreo. Estadística Española 31, 227-235.

|4| Vélez, R. (1977). Procesos Estocásticos, U.N.E.D., Madrid.

Rev. Academia de Ciencias. Zaragoza. 48 (1993)

INTERVALOS DE CONFIANZA BILATERALES PARA ESTIMAR UNA PROPORCIÓN

por

M. Ruiz Espejo

Departamento de Estadística

Facultad de Ciencias Económicas y Empresariales

Universidad Complutense de Madrid

Summary. The problem of bisided interval estimation of maximum absolute error of a proportion of a finite population under simple random sampling without replacement of effective size n (mas), it has not satisfactorily solved. We make a critical review of the usual methods and we propose an inferential method of minimum distance for building such intervals.

1. Introducción

El problema de estimación por intervalo de una proporción de población finita viene siendo tratado habitualmente en la práctica, suponiendo que la proporción muestral $\hat{p}$, cuando el tamaño efectivo fijo de la muestra es suficientemente grande, se distribuye como una variable aleatoria normal con media la proporción poblacional P, y varianza la del estimador proporción muestral $V(\hat{p})$.

En el caso en que el diseño es de muestreo aleatorio simple con reemplazamiento (masr) y el tamaño muestral n es suficientemente grande, se puede hacer uso del Teorema Central del Límite y suponer razonablemente que $\hat{p}$ se distribuye aproximadamente $N(P, \sqrt{PQ/n})$, siendo Q = 1-P, puesto que la proporción muestral $\hat{p}$ es una media aritmética de unos o ceros cada uno de los cuales se obtienen de variables aleatorias independientes e idénticamente distribuídas a la población, al seleccionar unidades con probabilidades iguales y con reemplazamiento antes de las sucesivas observaciones.

De este modo, llamando "e" al "error máximo de muestreo", tendremos que el intervalo $(\hat{p} \mp e)$ se obtiene igualando

$$e = k.\sigma(\hat{p}) \ , \qquad (1)$$

siendo k una constante obtenida de las tablas de la distribución normal para los diferentes niveles de confianza, por ejemplo,

k = 2 al nivel de confianza del 0.955, o bien,

k = 3 al nivel de confianza del 0.997.

De (1), elevando al cuadrado ambos miembros, tenemos

$$e^2 = k^2 V(\hat{p}) = k^2 PQ/n \ ,$$

por tratarse de un diseño de muestreo aleatorio simple con reemplazamiento (masr), de donde despejando n (tamaño muestral fijo), tendremos

$$n = k^2 PQ/e^2$$

ó

$$e = \sqrt{k^2 PQ/n} \leq \frac{1}{2}\sqrt{k^2/n}$$

en el caso más desfavorable en que $PQ \leq 1/4$.

Por un razonamiento similar, bajo muestreo aleatorio simple sin reemplazamiento (mas), partiríamos de que si N es el tamaño de la población finita,

$$e = k.\sigma(\hat{p})$$

o bien

$$e^2 = k^2 V(\hat{p}) = k^2 \frac{N-n}{N-1} \frac{PQ}{n} \ ,$$

es decir

$$n(N-1)e^2 = k^2 (N-n)PQ \ ,$$

o también

$$n[(N-1)e^2 + k^2 PQ] = k^2 NPQ \ ,$$

de donde el tamaño muestral efectivo sería

$$n = \frac{k^2 NPQ}{e^2(N-1) + k^2 PQ}$$

ó

$$e = \sqrt{\frac{k^2(N-n)PQ}{(N-1)n}} \leq \frac{1}{2}\sqrt{\frac{k^2(N-n)}{(N-1)n}} \quad , \qquad (2)$$

en el caso más desfavorable.

El único inconveniente es que al tratarse de un diseño mas, las observaciones sucesivas son sin reemplazamiento, por lo que son observaciones dependientes y por tanto no nos encontramos en las hipótesis tradicionales del Teorema Central del Límite, por lo que no hay garantías de que la constante k pueda ser obtenida de la distribución normal, y por lo tanto la solución (2) no es siempre válida desde este punto de vista.

Además, en base al trabajo de Plane y Gordon (1982), podemos asegurar que cuando el tamaño muestral n crece nunca se podrá decir que la distribución de la proporción muestral $\hat{p}$ sea normal bajo diseño mas.

Otros trabajos de interés sobre estimación por intervalo para proporciones son los de Buonaccorsi (1987), Katz (1953), Peskun (1990) y Wright (1991), que proporcionan soluciones exactas de intervalos unilaterales.

2. Una solución exhaustiva

La aproximación normal es válida con diseño masr de tamaño fijo n suficientemente grande; en los demás casos (diseño mas o cuando n es relativamente pequeño) una solución exhaustiva o por exceso se puede obtener haciendo uso de la desigualdad de Tchebycheff.

En este caso y en el visto en la sección 1, es interesante observar que dados e, N y $1-\alpha$, queda determinado n supuesto que se conoce P por la aproximación $\hat{p}$; pero también se podría partir de los datos n, N y $1-\alpha$, quedando determinado el error máximo e, cuando P esté suficientemente aproximado por $\hat{p}$ (El dato N es necesario con diseño mas, y no para el diseño masr).

Puesto que la proporción muestral $\hat{p}$ es insesgada para la proporción poblacional P y la varianza de $\hat{p}$ es conocida para los diseños masr y mas, podemos escribir que la probabilidad

$$p[|\hat{p} - P| < e] \geq 1 - \frac{V(\hat{p})}{e^2} = 1 - \alpha ,$$

siendo $V(\hat{p}) = PQ/n$ en masr y $V(\hat{p}) = (N-n)PQ/[(N-1)n]$ en mas. Al igualar $V(\hat{p})$ a $e^2\alpha$, despejando el tamaño muestral n, tendremos

$$n = \frac{PQ}{e^2\alpha} \qquad (3)$$

ó

$$e = \sqrt{\frac{PQ}{n\alpha}} \leq \frac{1}{2}\sqrt{\frac{1}{n\alpha}}$$

(en el caso más desfavorable) bajo diseño masr, y si igualamos también

$$V(\hat{p}) = \frac{N-n}{N-1}\frac{PQ}{n} = e^2\alpha ,$$

para diseño mas, resultando

$$n(N-1)e^2 = NPQ - nPQ .$$

Despejando n quedará

$$n[(N-1)e^2\alpha + PQ] = NPQ ,$$

es decir,

$$n = \frac{NPQ}{(N-1)e^2\alpha + PQ} \qquad (4)$$

ó

$$e = \sqrt{\frac{(N-n)PQ}{(N-1)n\alpha}} \leq \frac{1}{2}\sqrt{\frac{N-n}{(N-1)n\alpha}}$$

en el caso más desfavorable en que $PQ \leq 1/4$.

Las soluciones por exceso (3) y (4) dan siempre tamaños muestrales muy superiores a las obtenidas haciendo uso de tablas de la distribución normal.

En Cochran (1977) se sugieren algunas tablas estadísticas como las de las distribuciones binomial o hipergeométrica indicadas para los diseños masr y mas respectivamente, pero concretamente para este último diseño se aprecian

serias lagunas y limitaciones en la práctica, como es que N tenga que ser inferior o igual a 100, o valga 500, 2500 y 10000 pudiéndose interpolar gráficamente las soluciones para los valores intermedios de N, no tratándose en los casos de N superiores a 10000, como es muy frecuente en la práctica. Además no se indica ningún criterio para estimar de qué parámetros es la distribución binomial o hipergeométrica ya que sólo podrían conocerse sus parámetros con un diseño censal, y en este caso no sería necesario ya estimar por intervalo P, pues este parámetro sería conocido con exactitud.

3. Intervalos de confianza propuestos

En esta sección damos cobertura al problema no resuelto de asignar intervalos bilaterales de confianza de error máximo e para el diseño mas, siendo N cualquier número natural, no necesariamente pequeño (hasta N = 2000) como trata computacionalmente Wright (1991).

El estimador proporción muestral $\hat{p}$ bajo diseño mas, de tamaño efectivo fijo n, se distribuye

$$\hat{p} = X/n \ ,$$

de modo que X = H(N,n,NP) es una distribución hipergeométrica, es decir

$$p(X = k) = \frac{\binom{NP}{k}\binom{NQ}{n-k}}{\binom{N}{n}} \quad \text{si } 0 \leq k \leq \min\{n,NP\} \text{ y } 0 \leq n-k \leq \min\{n,NQ\} \ .$$

Dados ahora n, N y $1-\alpha$, el error máximo "e" se determina haciendo variar e = 0, 1/n, 2/n, ... hasta n/n = 1 de manera que verifique que el error máximo al nivel de confianza $1-\alpha$, "e", es el primer valor de e que satisface

$$\varphi(e - 1/n) < 1 - \alpha$$

y

$$\varphi(e) \geq 1 - \alpha \ ,$$

siendo

$$\varphi(e) = \sum_{x=n\hat{p}-ne}^{n\hat{p}+ne} p(X^* = x)$$

donde $X^* = H(N,n,\widehat{NP})$ es una estimación paramétrica de mínima distancia de X; es decir, $\widehat{NP}$ es un número natural estimado por redondeo que verifica

$$\widehat{NP} = \begin{cases} [N\hat{p}] & \text{si y sólo si } N\hat{p} - [N\hat{p}] \leq 1/2 \\ [N\hat{p}+1] & \text{si y sólo si } N\hat{p} - [N\hat{p}] > 1/2 \ . \end{cases} \qquad (5)$$

En (5) el símbolo "[...]" corresponde a "parte entera". Aunque la labor de determinar "e" es aparentemente complicada, este valor puede obtenerse directamente mediante un programa de ordenador con las ideas aquí sugeridas, de modo que el intervalo bilateral de confianza propuesto para P será

$$\widehat{NP}/N \mp e \ .$$

Referencias

|1| Buonaccorsi, J.P. (1987). A note on confidence intervals for proportions in finite populations. Amer. Statist. 41, 215-218.

|2| Cochran, W.G. (1977). Sampling Techniques (3ª edición), Wiley, Nueva York.

|3| Katz, L. (1953). Confidence intervals for the number showing a certain characteristic in a population when sampling is without replacement. J. Amer. Statist. Assoc. 48, 256-261.

|4| Peskun, P.H. (1990). A note on a general method for obtaining confidence intervals from samples from discrete distributions. Amer. Statist. 44, 31-35.

|5| Plane, D.R. y Gordon, K.R. (1982). A simple proof of the nonapplicability of the central limit theorem to finite populations. Amer. Statist. 36, 175-176.

|6| Wright, T. (1991). Exact Confidence Bounds when Sampling from Small Finite Universes, Springer-Verlag, Nueva York.

Rev. Academia de Ciencias. Zaragoza 48 (1993)

SOBRE LA INFERENCIA CON DATOS MUESTRALES PARA ESTUDIOS ANALITICOS

por

M. Ruiz Espejo y A. Arcos Cebrián

Departamento de Estadística
Facultad de Ciencias Económicas y Empresariales
Universidad Complutense
28223 Madrid

y

Departamento de Estadística
Facultad de Ciencias
Universidad de Granada
18071 Granada

Summary. In some cases a fixed finite population, of size N, may be considered in itself a simple random sample (independent observations among them obtained from an infinite random variable Y). The interest is based usually to infer on infinite population mean, variance or variance estimation. In the present paper we present the simplest strategies for this purpose, and consequently of the biggest practical interest.

1. Introducción

El modelo superpoblacional más sencillo en cierto sentido (ver Cassel, et al. |1|) es considerar que el valor y_i (i=1,2,...,N) de la variable de interés "y", es una observación independiente que procede de la realización de una variable aleatoria Y_i (i=1,2,...,N) que es independiente e idénticamente distribuída a la variable aleatoria Y de la que pretendemos inferir sobre su media poblacional $\bar{\mu} = E(Y)$, y su varianza poblacional $\bar{\sigma}^2 = V(Y)$.

Este planteamiento general ha sido tratado por Koop |3| para el estudio de ciertos problemas bajo este modelo. Una de sus consecuencias es que para cualquier diseño muestral no informativo, la media aritmética basada

en el tamaño muestral efectivo fijo, es un estimador UMV (uniformemente de mínima varianza) de $\bar{\mu}$, media poblacional de la variable aleatoria Y. Otros estudios recientes de muestreo estratificado en estudios analíticos han sido dados por Ruiz |4|.

Así observamos que el problema de inferencia sobre $\bar{\mu}$ ó $\bar{\sigma}^2$ admite dos tipos de aleatorización; una basada en el modelo M propuesto por el cual es posible observar la variable de interés en una población finita U de tamaño N, y otra basada en el diseño muestral (ordenado o no) por el que se selecciona la muestra de tamaño fijo n (≥ 1).

La primera fase de aleatorización basada en el modelo M proporciona el universo finito existente con sus N posibles observaciones; la segunda fase de aleatorización se basa en un diseño muestral que es controlable por el investigador estadístico.

Por tanto es muy importante estar cerciorado de que los posibles valores observables y_i (i=1,2,...,N) procedan todos de la misma población estadística Y, y además que sean independientes entre sí las N posibles observaciones asociadas a las N unidades identificables que constituyen la población o universo finito U. Si no fuera así, la hipótesis hecha sobre el modelo M no sería razonable y el uso de las técnicas que presentamos a continuación no serían apropiadas en la práctica.

A pesar del resultado comentado de Koop |3| sobre estimación UMV, el uso de diseños ordenados pueden ser ventajosos por su costo esperado inferior al de los diseños no ordenados, para un tamaño muestral fijo n. Esta afirmación puede ser argumentada para estudios analíticos del mismo modo que en el contexto de población finita fijada (Ruiz y Santos, |5|).

2. Estrategias insesgadas para la media

Proposición 1. Los estimadores tradicionales "media muestral" $\bar{y}_s$ ó $\bar{y}_{\underline{s}}$ son insesgados para estimar $\bar{\mu}$, bajo el modelo M y para diseños de muestreo aleatorio simple sin o con reemplazamiento respectivamente (mas ó masr)□

Demostración

$$E(\bar{y}_s) = E_M E_{mas}(\bar{y}_s) = E_M(\bar{y}) = \bar{\mu} ,$$

pues

$$E_{mas}(\bar{y}_s) = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y}$$

(ver Cassel, et al. |1|), y

$$E_M(\bar{y}) = \frac{1}{N}\sum_{i=1}^{N} E_M(y_i) = \frac{1}{N} N\bar{\mu} = \bar{\mu}$$

por estar Y_i distribuída como Y. Es decir, $E(mas\text{-}M,\bar{y}_s) = \bar{\mu}$. También

$$E(\bar{y}_{\underline{s}}) = E_M E_{masr}(\bar{y}_{\underline{s}}) = E_M(\bar{y}) = \bar{\mu} ,$$

sabiendo que la estrategia $(masr,\bar{y}_{\underline{s}})$ es insesgada para $\bar{y}$ en el modelo de población finita fijada. Así, $E(masr\text{-}M,\bar{y}_{\underline{s}}) = \bar{\mu}$ □

Con este primer resultado concluímos que $(mas\text{-}M,\bar{y}_s)$ y $(masr\text{-}M,\bar{y}_{\underline{s}})$ son estrategias (diseño-modelo, estimador) insesgadas para estimar $\bar{\mu}$. Las varianzas para estas estrategias son dadas a continuación (supuesto que la varianza $\bar{\sigma}^2 < \infty$).

Proposición 2. Siendo n el tamaño muestral fijo,

$$V(mas\text{-}M,\bar{y}_s) = \frac{1}{n}\bar{\sigma}^2 \qquad \text{y} \qquad V(masr\text{-}M,\bar{y}_{\underline{s}}) = \frac{N+n-1}{Nn}\bar{\sigma}^2 \quad \square$$

Demostración

Es consecuencia del uso del teorema de Madow en ambos casos,

$$V(mas\text{-}M,\bar{y}_s) = E_M V_{mas}(\bar{y}_s) + V_M E_{mas}(\bar{y}_s) = E_M(\frac{N-n}{Nn} S^2) + V_M(\bar{y}) =$$

$$= \frac{N-n}{Nn}\bar{\sigma}^2 + \frac{1}{N}\bar{\sigma}^2 = \frac{1}{n}\bar{\sigma}^2$$

(ver Cochran, |2|), siendo S^2 la cuasivarianza de la población finita,

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N} (y_i - \bar{y})^2 .$$

$$V(masr\text{-}M,\bar{y}_{\underline{s}}) = E_M V_{masr}(\bar{y}_{\underline{s}}) + V_M E_{masr}(\bar{y}_{\underline{s}}) = E_M(\frac{N-1}{Nn} S^2) + V_M(\bar{y}) =$$

$$= \frac{N-1}{Nn}\bar{\sigma}^2 + \frac{1}{N}\bar{\sigma}^2 = \frac{N+n-1}{Nn}\bar{\sigma}^2 \quad \square$$

Obviamente, comparando ambas estrategias, vemos que es más precisa la primera (mas-M,$\bar{y}_s$) que la segunda (masr-M,$\bar{y}_{\underline{s}}$) siempre que $n>1$. Sin embargo, la segunda es más económica en promedio que la primera para un tamaño muestral fijo común n (>1) según se puede ver en Ruiz y Santos |5|. Es interesante observar que las varianzas obtenidas para estudios analíticos son distintas de las clásicas V(mas,$\bar{y}_s$) y V(masr,$\bar{y}_{\underline{s}}$) en poblaciones finitas, aunque conservan sus propiedades comparativas (el muestreo sin reemplazamiento es más eficiente que con reemplazamiento, pero de mayor coste esperado).

Denotando por (p-M,t_I) a las estrategias intermedias (introducidas por Ruiz y Santos |5|) para estudios analíticos, tenemos que si

$$t_I = \frac{1}{m}\sum_{h=1}^{m} \bar{y}_h ,$$

Proposición 3. Siendo N el tamaño poblacional y n el tamaño de la muestra independiente h (=1,2,...,m),

$$E(p\text{-}M,t_I) = \bar{\mu} \qquad \text{y} \qquad V(p\text{-}M,t_I) = \frac{N-n+mn}{mnN}\bar{\sigma}^2 \quad \square$$

3. Estrategias insesgadas para la varianza

Si suponemos que $\bar{\sigma}^2 = V(Y) < \infty$, tenemos el siguiente resultado

Proposición 4. Dos estrategias insesgadas para $\bar{\sigma}^2$ son (mas-M,s_s^2) y (masr-M, $\frac{N}{N-1}\, s_{\underline{s}}^2$), siendo s^2 la cuasivarianza muestral para la muestra s ó $\underline{s}$ que se subindica $\square$

Demostración

$$E(mas\text{-}M,s_s^2) = E_M E_{mas}(s_s^2) = E_M(S^2) = \bar{\sigma}^2 ,$$

siendo

$$s_s^2 = \frac{1}{n-1}\sum_{i\in s}(y_i - \bar{y}_s)^2 .$$

También,

$$E(masr\text{-}M,\frac{N}{N-1}\, s_{\underline{s}}^2) = E_M E_{masr}(\frac{N}{N-1}\, s_{\underline{s}}^2) = E_M(\frac{N}{N-1}E_{masr}(s_{\underline{s}}^2)) =$$

$$= \frac{N}{N-1} E_M(\sigma^2) = \frac{N}{N-1} \frac{N-1}{N} \bar{\sigma}^2 = \bar{\sigma}^2 ,$$

siendo

$$s^2_{\underline{s}} = \frac{1}{n-1} \sum_{i \in \underline{s}} (y_i - \bar{y}_{\underline{s}})^2 \qquad y \qquad \sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y})^2$$

la cuasivarianza muestral ordenada y la varianza de la población finita respectivamente □

Corolario 1. Estimadores insesgados de las varianzas dadas en la proposición 2 son

$$\hat{V}(\text{mas-M}, \bar{y}_s) = \frac{1}{n} s^2_s \qquad y \qquad \hat{V}(\text{masr-M}, \bar{y}_{\underline{s}}) = \frac{N+n-1}{n(N-1)} s^2_{\underline{s}} \quad \square$$

Demostración

Es inmediata a partir de las proposiciones 2 y 4 □

Corolario 2. Un estimador insesgado de la varianza $V(p\text{-}M, t_I)$ (dada en la proposición 3), es

$$\hat{V}(p\text{-}M, t_I) = \frac{N-n+mn}{m^2 nN} \sum_{h=1}^{m} s^2_h ,$$

siendo s^2_h la cuasivarianza muestral en la muestra independiente $h(=1,2,\ldots,m)$ □

Referencias

(1) Cassel, C.M., Särndal, C.E. y Wretman, J.H. (1977). *Foundations of Inference in Survey Sampling*. Wiley. Nueva York.

(2) Cochran, W.G. (1977). *Sampling Techniques* (3ª edición). Wiley. Nueva York

(3) Koop, J.C. (1986). Some problems of statistical inference from sample survey data for analytic studies. *Statistics* **17**, 237-247.

(4) Ruiz, M. (1991). Muestreo estratificado en estudios analíticos. *Metron* **49** 459-468.

(5) Ruiz, M. y Santos, J. (1989). Estrategias intermedias de muestreo. *Estadística Española* **31**, nº 121, 227-235.

Rev. Academia de Ciencias. Zaragoza. 48 (1993)

THE PROBLEM OF OPTIMUM WEIGHTS IN STRATIFIED SAMPLING

by

M. Ruiz Espejo* and M. Rueda García**

*Departamento de Estadística
Facultad de Ciencias Económicas y Empresariales
Universidad Complutense de Madrid

and

**Departamento de Estadística
Facultad de Ciencias Matemáticas
Universidad de Granada

Summary. In this paper we are studying the gain in precision of stratified sampling, being possible for any stratification and any sample allocation, considering variable the weights of the usual estimator of the population mean and optimizing such weights. As a result of this we can build new theoretical unbiased estimators with lesser variance than the classical stratified estimators of equal sample size, as we can see in an example included in section 3.

1. Introduction

Stratified sampling is a traditional way to estimate the mean of a finite population, which may have certain advantages due to its higher precision than some classical estimators as for example the sample mean in some cases (Cochran, 1953). This paper considers an important theoretical question, if the precision of the ordinary stratified sampling estimator can be improved by optimizing the weights of the strata sample means. The answer is that it can.

Some steps forward have been given in order to improve the precision of the classical stratified estimator of the population mean, among them the theoretical problem of sample allocation or distribution of the sample size among the strata whose optimization is due to Tschuprow-Neyman, or also the optimization of strata building or the theoretical problem of optimum strat-

ification studied by Dalenius (see Cochran, 1977); so, in this paper we introduce and discuss the problem of optimizing the weights being used in the traditional estimator

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h$$

where "y" is the interest variable, L the number of strata, $W_h = P_h = N_h/N$ (h=1,2,...,L) the relative size of stratum h, and $\bar{y}_h$ the sample mean estimator in the stratum h (=1,2,...,L). Though W_h is usually fixed and equal to N_h/N, being N_h the size of stratum h, and N the size of the whole finite population

$$N = \sum_{h=1}^{L} N_h \ ,$$

in order to improve greatly the precision of the stratified sampling we may consider W_h variable so that $\bar{y}_{st}$ is unbiased for the population mean μ. This new optimization is compatible with the ones already known; being given a stratification and an allocation (which may be optimum), we optimize in turn the weights W_h (h=1,2,...,L).

2. The problem of optimization

The variance of the estimator $\bar{y}_{st}$, $V(\bar{y}_{st})$, will be the function to minimize

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 V(\bar{y}_h)$$

being subject to unbiased constraint

$$\mu = \sum_{h=1}^{L} W_h \mu_h \quad [= E(\bar{y}_{st})]$$

being μ_h $[= E(\bar{y}_h)]$ the mean of the interest variable "y", in the stratum h (=1,2,...,L). Using the Lagrangian (where λ is the Lagrange multiplier)

$$\mathcal{L}^* = \sum_{h=1}^{L} W_h^2 V(\bar{y}_h) - \lambda \left(\sum_{h=1}^{L} W_h \mu_h - \mu \right)$$

which we solve as follows

$$\frac{\partial \mathcal{L}^*}{\partial W_h} = 2W_h V(\bar{y}_h) - \lambda \mu_h = 0 \implies W_h = P_h^* = \frac{\lambda \mu_h}{2V(\bar{y}_h)} \ . \qquad (1)$$

As,

$$\mu = \sum_{h=1}^{L} P_h^* \mu_h = \frac{\lambda}{2} \sum_{h=1}^{L} \frac{\mu_h^2}{V(\bar{y}_h)} \Longrightarrow \quad \lambda = \frac{2\mu}{S_2}, \tag{2}$$

being

$$S_i = \sum_{h=1}^{L} \mu_h^i / V(\bar{y}_h) \ , \quad i=0,1,2.$$

Thus (1) and (2) imply

$$W_h = P_h^* = \frac{\mu \mu_h}{V(\bar{y}_h) S_2} \ , \tag{3}$$

being these the optimum weights that minimize $V(\bar{y}_{st})$, where $\bar{y}_{st}$ is unbiased for the population mean with such weights. This procedure is simple and valid in the case of $L \geq 2$ (as there is only one constraint), for any stratification, for any allocation and for any unbiased estimator $\bar{y}_h$ (for μ_h) used in the stratum h.

The minimum variance will be, in this case

$$V_{opt}(\bar{y}_{st}^*) = \sum_{h=1}^{L} P_h^{*2} V(\bar{y}_h) = \frac{\mu^2}{S_2} \ .$$

In stratified random sampling, the estimator of μ is

$$\bar{y}_{st}^* = \sum_{h=1}^{L} P_h^* \bar{y}_h$$

and its variance will be

$$V_{opt}(\bar{y}_{st}^*) = \mu^2 / [\sum_{h=1}^{L} \mu_h^2 / (\frac{N_h - n_h}{N_h - 1} \cdot \frac{\sigma_h^2}{n_h})]$$

where σ_h^2 is the variance of stratum h.

In this optimum case, the sum

$$\sum_{h=1}^{L} P_h^*$$

does not have to be 1. If we wanted to include a new unnecessary constraint,

$$1 = \sum_{h=1}^{L} W_h ,$$

the problem of optimization that may arise would be

$$\mathcal{L} = \sum_{h=1}^{L} W_h^2 V(\bar{y}_h) - \lambda_1 \left(\sum_{h=1}^{L} W_h \mu_h - \mu \right) - \lambda_2 \left(\sum_{h=1}^{L} W_h - 1 \right)$$

(being λ_1 and λ_2 the Lagrange multipliers), which would be solved as follows

$$W_h = P'_h = \frac{\lambda_1 \mu_h + \lambda_2}{2V(\bar{y}_h)}$$

with

$$\lambda_2 = \frac{2(\mu S_1 - S_2)}{S_1^2 - S_0 S_2} \qquad \text{and} \qquad \lambda_1 = \frac{2 - \lambda_2 S_0}{S_1} .$$

As there are two constraints now, the gain in precision happens to be effective for $L \geq 3$ although this gain is in any case lower than the one given by (3).

3. Example

Being the given data, L=3,

h	1	2	3	
μ_h	1	4	10	
N_h	6	9	3	
n_h	2	3	1	(Proportional allocation)

h	1	2	3
σ_h^2	1	2	3
P_h	1/3	1/2	1/6
P_h^*	0.1474202	0.4717445	0.1965602
P'_h	0.3296703	0.5054945	0.1648352

$$\sum_{h=1}^{3} P_h^* = 0.8157249 \; ; \quad \sum_{h=1}^{3} P_h' = 1 \; ;$$

$$\sum_{h=1}^{3} P_h^* \mu_h = \sum_{h=1}^{3} P_h' \mu_h = \mu = 4 \; .$$

Taking $\bar{y}_h$ as the sample mean with "simple random sampling without replacement" design, being these independent among the strata. The relative gains in precision with respect to the classical procedure (whose variance is $V(\bar{y}_{st})$) are

$$\frac{V(\bar{y}_{st})}{V_{opt}(\bar{y}^*_{st})} = \frac{0.2527778}{0.2358723} = 1.071672 > 1$$

using

$$\bar{y}^*_{st} = \sum_{h=1}^{3} P_h^* \bar{y}_h \qquad \text{for} \qquad V_{opt}(\bar{y}^*_{st}) \; ;$$

and

$$\frac{V(\bar{y}_{st})}{V_{opt}(\bar{y}'_{st})} = \frac{0.2527778}{0.2527473} = 1.000121 > 1$$

using

$$\bar{y}'_{st} = \sum_{h=1}^{3} P_h' \bar{y}_h \qquad \text{for} \qquad V_{opt}(\bar{y}'_{st}) \; ;$$

and where

$$\bar{y}_{st} = \sum_{h=1}^{3} P_h \bar{y}_h \qquad \text{for} \qquad V(\bar{y}_{st}) \; .$$

This confirms the previous statements made in the aforementioned sections as the relative gain in precision in the example, is superior to 7 %, concretely 7.1672 %. As a result of this, the proposed technique is considered quite useful in theoretical conventional stratified sampling. We can consider some cases in which if μ_L is relatively large and $V(\bar{y}_L)$ is relatively small, the gains in precision or accuracy are greater than in the example proposed in this section.

4. A general justification

For any case, the ratio $R = V(\bar{y}_{st})/V_{opt}(\bar{y}^*_{st})$ will be

$$R = \frac{V(\bar{y}_{st})}{V_{opt}(\bar{y}^*_{st})} = \frac{\frac{1}{N^2}\sum_{h=1}^{L} N_h^2 V(\bar{y}_h)}{\frac{1}{N^2}\frac{(\sum_{h=1}^{L} N_h \mu_h)^2}{\sum_{h=1}^{L} \frac{\mu_h^2}{V(\bar{y}_h)}}} = \frac{[\sum_{h=1}^{L} N_h^2 \sigma^2(\bar{y}_h)][\sum_{h=1}^{L} \frac{\mu_h^2}{\sigma^2(\bar{y}_h)}]}{(\sum_{h=1}^{L} N_h \mu_h)^2} ,$$

and from Cauchy-Schwarz inequality,

$$R \geqslant \frac{(\sum_{h=1}^{L} N_h \mu_h)^2}{(\sum_{h=1}^{L} N_h \mu_h)^2} = 1 .$$

Or also, $V_{opt}(\bar{y}^*_{st}) \leqslant V(\bar{y}_{st})$ in the most general case.

References

|1| Cochran, W.G. (1953). Sampling Techniques (1st. edition), Wiley, New York.

|2| Cochran, W.G. (1977). Sampling Techniques (3rd. edition), Wiley, New York.

Rev. Academia de Ciencias. Zaragoza. 48 (1993)

A MODIFIED JELLIUM MODEL FOR SMALL METAL CLUSTERS

F. Castaño[1], M. Membrado[2], A.F. Pacheco[2,3] and J. Sañudo[1]

[1] *Departamento de Física. Facultad de Ciencias. Universidad de Extremadura. 06071 Badajoz. (Spain).*
[2] *Departamento de Física Teórica. Facultad de Ciencias . Universidad de Zaragoza. 50009 Zaragoza. (Spain).*
[3] *ICMA (CSIC). Facultad de Ciencias. Universidad de Zaragoza. 50009 Zaragoza. (Spain).*

ABSTRACT

Using the Density-Functional theory for the free electrons of a metal cluster we have studied the validity of the jellium model to describe its ion background. Depending on the metal and the number of atoms in the cluster, a hollow uniform-density sphere is proposed as a more correct model.

1. INTRODUCTION

In recent years, a lot of research has been devoted to analysing the physics of small metal clusters[1,2]. This is due to both their intrinsic scientific interest and their enormous technological projections. Because of their smallness, the structure of these objects is akin to that of metallic surfaces and hence, these systems constitute a paradigmatic object where the transition of any physical properties from the atom to the bulk material can be analysed. Among the different theoretical strategies used to analyse the free-electron structure of clusters, the Density-Functional theory[3], in its different versions[4], is one of the most popular. With respect to the ion background, the jellium model is the most habitual because of its simplicity, universality, and correctness of many of its predictions[2,5,6]. Unfortunately, the jellium approximation poses problems in several issues of metal physics and in particular, when it is used to describe the surface of dense metals[7], which is why one could easily foresee the appearance of difficulties when applying the jellium to cluster physics. [For the bulk metal, several problems derived from the jellium are solved when the effects of the pseudopotential are adequately

included; in this respect see the recent work contained in Reference 8]. Thus, it is interesting to try to find out, from first principles, when the jellium approximation is valid for clusters and, in the negative cases, formulate simple alternatives. In a recent paper[9], which we will refer to as I, we have analyzed the same problem through a simple variational model (also used successfully in the description of non metal[10] clusters) based on a trial function for the electron density. The results emerging from I have spurred us to carry out this second work which basically confirms the conclusions of I. Now we use again the Density-Functional theory but, by solving exactly the Euler-Lagrange equations of the system.

As our aim is to show with total clarity when the jellium approximation is valid (or not) for small metal clusters, we develop a simple model which, including the jellium as a particular case, describes in general a uniform-background model allowing a concentric hole to appear in its interior. Thus the global spherical symmetry of the problem is maintained. The tendency of a cluster to develop such a hole, i.e. to increase its surface, is interpreted as the onset of the instability of the jellium which could preclude its complete breakdown.

2. THE MODEL

The total energy of the cluster is described by the following functional:

$$E[n_e] = \frac{1}{2}\int \phi(\boldsymbol{r})\left[n^+(\boldsymbol{r}) - n_e(\boldsymbol{r})\right] d\boldsymbol{r} + G[n_e], \tag{2.1}$$

where n^+, the constant charge density of the positive distribution, is taken as that of the bulk material, n_e is the electron charge density, ϕ is the total electrostatic potential, and G is chosen as

$$G[n_e] = \frac{3}{10}(3\pi^2)^{2/3}\int n_e^{5/3} d\boldsymbol{r} - \frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}\int n_e^{4/3} d\boldsymbol{r}$$
$$- 0.056\int \frac{n_e^{4/3}}{0.079 + n_e^{1/3}} d\boldsymbol{r} + \frac{1}{72}\int \frac{(\nabla n_e)^2}{n_e} d\boldsymbol{r}. \tag{2.2}$$

Thus, $E[n_e]$ contains the kinetic term (plus its first gradient correction), the electrostatic, exchange, and correlation energies. The coefficient of the gradient term is what is derived from the unambiguous Kirzhnits expansion[4a], and the correlation is of the Wigner type. In the previous formulas and throughout the paper Hartree units (a.u.), $\hbar = m = e = 1$, are used. The electrostatic potential is linked to the charge density through the Poisson equation:

$$\nabla^2\phi = \nabla^2(\phi_e + \phi^+) = 4\pi(n_e - n^+), \tag{2.3}$$

where ϕ_e and ϕ^+ are the potentials created by the electrons and positive background respectively. The cluster energy resulting from Eq.(1) must be an extremum with respect to density variations subject to the prescribed normalization. Thus we must solve:

$$\frac{\delta}{\delta n_e}\left\{E[n_e] - \mu\int n_e d\boldsymbol{r}\right\} = 0, \tag{2.4}$$

where μ is the chemical potential. Eq.(2.4) leads us to the Euler-Lagrange equations of this system. This is the difference with respect to I: there we used a trial function for

the density, whilst here we look for the exact minimum of Eq.(2.1) (i.e. the solution of Eq.(2.4)). $R(H)$ is the exterior (interior) radius of the positive core $n^+(\boldsymbol{r}) = n_0\theta(R - r)\theta(r - H)$; N is the number of atoms and v the valence of the material. Thus we have $\frac{4\pi}{3}(R^3 - H^3)n_0 = Nv = Z$, and $n_0 = \frac{3}{4\pi r_s^3}$, r_s is the Wigner-Seitz parameter of the metal.

In our model the-analytic-expression for ϕ^+ is given by

$$\phi^+ = \begin{cases} \frac{3Z}{2R}\frac{(1-\lambda^2)}{(1-\lambda^3)}, & r \leq H; \\ \frac{Z}{2R^3(1-\lambda^3)}\left[3R^2 - r^2 - \frac{2\lambda^3R^3}{r}\right], & H < r < R; \\ \frac{Z}{r}, & r \geq R; \end{cases} \tag{2.5}$$

where the dimensionless parameter $\lambda = \frac{H}{R}$ has been used. Thus, inserting Eqs.(2.1), (2.3), and (2.5) into (2.4), we express the Euler-Lagrange equations in terms of the variable n_e.

3. RESULTS AND DISCUSSION

Our results are summarized in Fig.1, which shows the value of λ for four cases $N = 2, 20, 92$ and 186 in different materials. As a general tendency we observe how λ decreases as N increases. There is a set of elements, Na, K, etc. for which even with $N = 2$, λ is null and the jellium model is OK. At the other extreme: Cu, Mg, etc., we see that even for $N = 186$, the solid jellium fails, and our model is more appropriate. We also observe a zone of transition, where each N fixes a different value of λ. On occasions, there are "sudden falls" as is the case with La and also with Ag and Au. It seems that when the fall occurs for λ next to 1, it is abrupt. In copper, where for $N = 186$ clusters behave as pure surfaces, the fall starts for bigger systems, and thus for $N = 1860$ $\lambda \approx 0.17$. Our results, which are shown in Fig.1, follow the general trend of those obtained in paper I.

In Fig.2 we depict for gold clusters the explicit form adopted in our model by the electron density distributed around the positive background. This metal was chosen as being characteristic in the area of intermediate lambdas.

Thus, we conclude that the jellium approximation, which is correct for all alcaline metal clusters, improves, in general, with the increase in N[11]. These results suggest that our hollow-jellium-model, which includes the ordinary uniform sphere as a particular case, may be more suitable in those cases where the pure jellium model fails. Consequently, a number of calculations, of Khon-Sham type for example, would be worth redoing to allow the appearance of the inner hole; in all probability, this would considerably improve many theoretical predictions. In addition, we will say that corrections to spherity[12] could be easily included as a mere refinement of this model.

ACKNOWLEDGEMENTS

Helpful discussions held with M. Barranco, M. Pi, J.A. Alonso, and L.C. Balbás are gratefully acknowledged. We also want to thank the group of Barcelona for letting us using their variational codes. This work was supported by the *DGICYT* (PB90-0916).

REFERENCES

[1] See, for example, W.P. Halperin, Rev. Mod. Phys. 58, 533 (1986), and references therein.

[2] W.A. de Heer, W.D. Knight, M.Y. Chou, and M.L. Cohen, Solid State Phys. 40, 93 (1987), and references therein.

[3] P. Hohenberg and W. Kohn, Phys. Rev. 136, B864 (1964).

[4] a) See, for example, *Density Functional Methods in Physics* (Edited by R.M. Dreizler and J. da Providencia), Plenum, New York (1985), and references therein; D.R. Snider and R.S. Sorbello, Solid State Comm. 47, 845 (1983).

[5] W. Ekardt, Phys. Rev. *B*29, 1558 (1984); D.E. Beck, Solid State Comm. 49, 381 (1984); G. Makov, A. Nitzan and L.E. Brus, J. Chem. Phys. 88, 5076 (1988).

[6] M. Membrado, A.F. Pacheco, and J. Sañudo, Phys. Rev. *B*41, 5643 (1990); A. Mañanes, M. Membrado, J. Sañudo, A.F. Pacheco, and L.C. Balbás, Z. Phys. *D*19, 55 (1991).

[7] N.D. Lang, in *Solid State Physics* (Edited by F. Seitz, D. Turnbull and H. Ehrenreich), vol. 28, p. 225, Academic, New York (1973) and references therein.

[8] J.P. Perdew, H.Q. Tran, and E.D. Smith, Phys. Rev. *B*42, 11627 (1990).

[9] M. Membrado, A.F. Pacheco, and J. Sañudo, Solid State Comm. 77, 887 (1991).

[10] L.C. Balbás, A. Mañanes, M. Membrado, A.F. Pacheco, and J. Sañudo, J. Chem. Phys. 94, 7335 (1991).

[11] W.A. de Heer, P. Milani, and A. Châtelain, Phys. Rev. Lett. 63, 2834 (1989).

[12] K. Clemenger, Phys. Rev. *B*32, 1359 (1985).

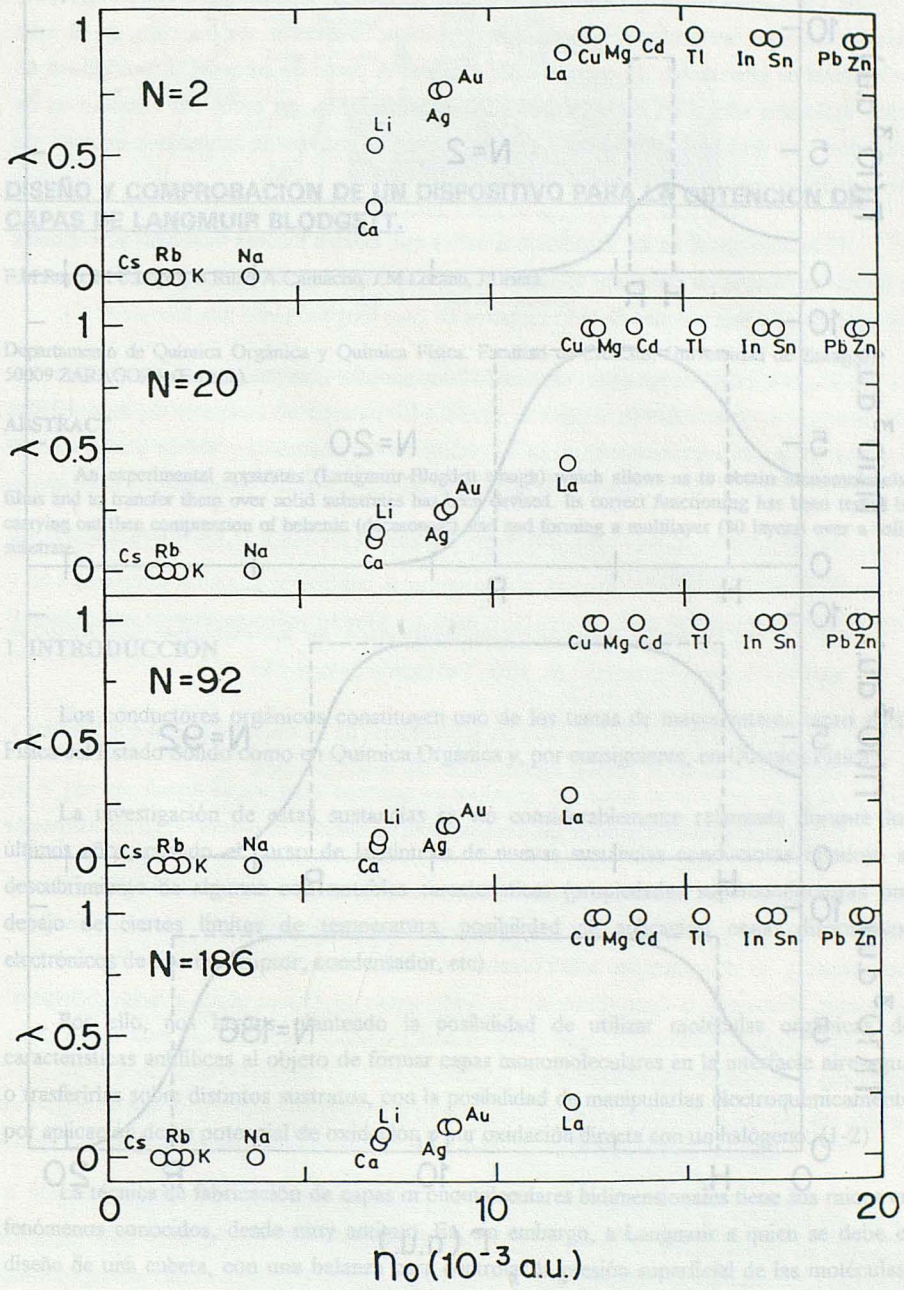


Figure 1. Value of λ for several metal clusters; the comparison among $N = 2, 20, 92$ and 186 illustrates its systematic shortening as N incrases.

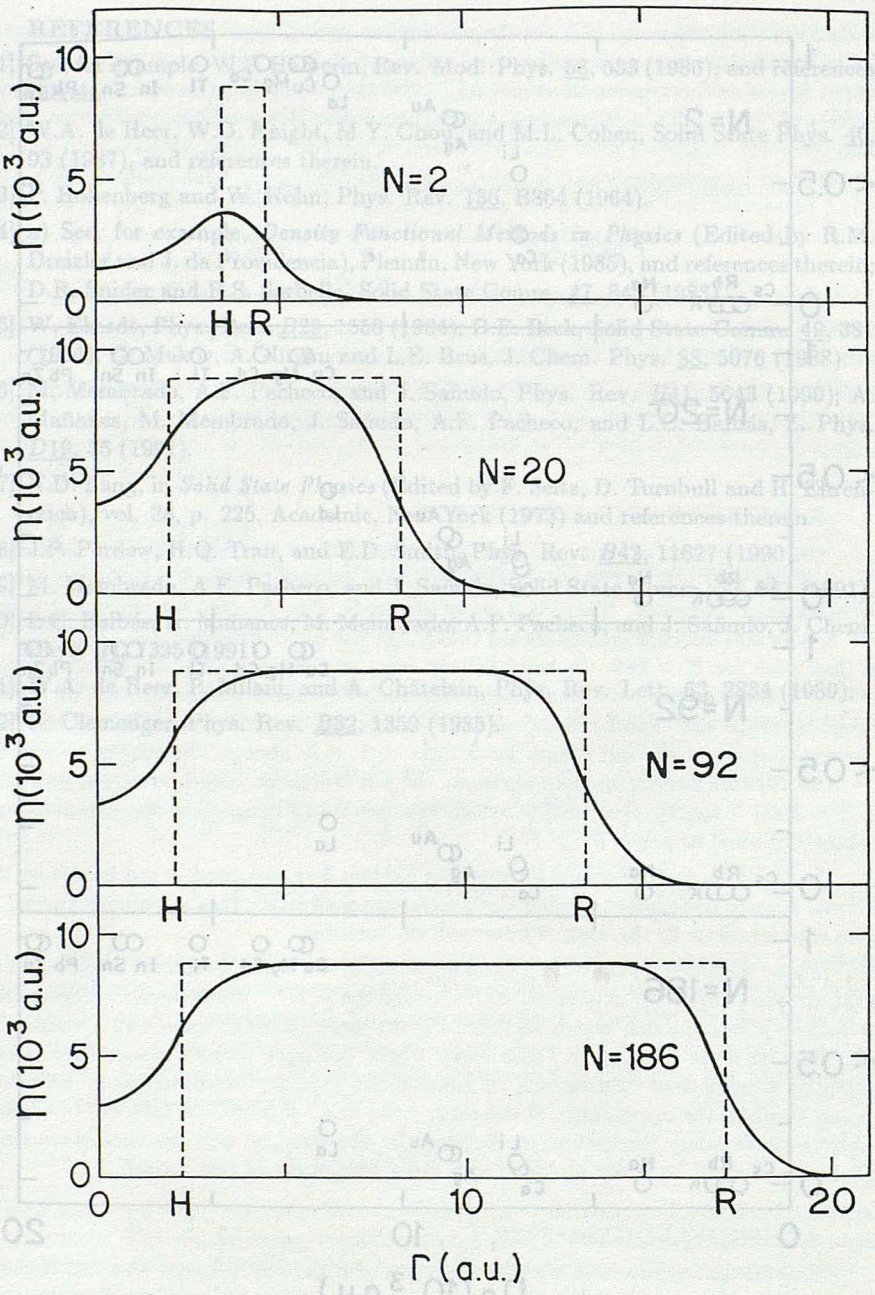


Figure 2. Electron distribution in gold clusters. The dashed line depicts the positive core with exterior and interior radii, R and H, respectively.

Rev. Academia de Ciencias. Zaragoza. 48 (1993)

DISEÑO Y COMPROBACION DE UN DISPOSITIVO PARA LA OBTENCION DE CAPAS DE LANGMUIR BLODGETT.

F.M.Royo, M.C.López, B.Ruiz, A.Camacho, J.M.Lozano, J.Urieta.

Departamento de Química Orgánica y Química Física. Facultad de Ciencias, Universidad de Zaragoza. 50009 ZARAGOZA (España).

ABSTRACT

An experimental apparatus (Langmuir-Blogdett trough) which allows us to obtain monomolecular films and to transfer them over solid substrates has been devised. Its correct functioning has been tested by carrying out then compression of behenic (docasonoic) acid and forming a multilayer (30 layers) over a solid substrate.

1. INTRODUCCION

Los conductores orgánicos constituyen uno de los temas de mayor interés tanto en la Física del Estado Sólido como en Química Orgánica y, por consiguiente, en Química Física.

La investigación de estas sustancias se vió considerablemente relanzada durante los últimos años, cuando el curso de la síntesis de nuevas sustancias conductoras condujo al descubrimiento de algunas con notables características (propiedades superconductoras por debajo de ciertos límites de temperatura, posibilidad de aplicación como dispositivos electrónicos de tipo interruptor, condensador, etc).

Por ello, nos hemos planteado la posibilidad de utilizar moléculas orgánicas de características anfifílicas al objeto de formar capas monomoleculares en la interfacie aire-agua o trasferirlas sobre distintos sustratos, con la posibilidad de manipularlas electroquímicamente por aplicación de un potencial de oxidación o por oxidación directa con un halógeno. (1-2)

La técnica de fabricación de capas m onomoleculares bidimensionales tiene sus raices en fenómenos conocidos, desde muy antiguo. Es, sin embargo, a Langmuir a quien se debe el diseño de una cubeta, con una balanza para controlar la presión superficial de las moléculas, que permite obtener capas monomoleculares, y él es también el primero en transferir una capa de ese tipo desde la superficie del agua a un sustrato sólido en 1919 (3).

Posteriormente, en 1935, (4) Katharine Blodgett muestra la manera en que se pueden trasferir de forma secuenciada las capas monomoléculares sucesivamente sobre un sustrato. La estructura laminar obtenida se conoce universalmente bajo el nombre de capas de Langmuir-Blodgett.

En 1962 Hans Kuhn (5) sintetiza y utiliza moléculas especialmente concebidas para construir edificios moleculares organizados y activos, centrando sus estudios en el área fundamental de intercambio de energía entre átomos. A partir de entonces, la posibilidad de incluir moléculas activas en edificios moleculares organizados, así como sus espectativas de utilización en diversas áreas científicas, promueven el interés de numerosos equipos de investigación en este campo.

El balance actual de los estudios nos indica que existen muchas moléculas susceptibles de formar multicapas de Langmuir Blodgett (LB), que son, en general, compuestos orgánicos que poseen una o más cadenas de hidrocarburos de gran longitud (16 o más átomos de C).

Se abre, por consiguiente, un campo muy amplio que puede extenderse en varias direcciones.

(a) Síntesis de las moléculas susceptibles de formar multicapas.

(b) Fabricación de las capas con la posibilidad de modificar la estructura a través de métodos electroquímicos.

(c) Estudio de las propiedades y caracterización de las multicapas por métodos espectrocópicos (por ej: UV, IR, RPE, Difracción de rayos X).

(d) Estudio de la correlación entre las propiedades de las multicapas y las condiciones y naturaleza tanto de la subfase líquida como del sustrato sólido.

(e) Aplicación tecnológica de las multicapas.

Al objeto de contribuir a este novedoso e interesante tema de investigación se ha puesto en marcha, la técnica experimental (cuba de Langmuir) que permita la obtención de las monocapas y su transferencia sobre sustratos sólidos y se ha comprobado su correcta utilización efectuando la compresión de un ácido graso de cadena larga: el ácido behénico. Finalmente, se han trasferido 30 capas sobre un sustrato sólido formando una multicapa y se ha verificado la bondad de la transferencia.

2.DISPOSITIVO EXPERIMENTAL

La cubeta, cuyas dimensiones son 370x149x9 mm excepto en la zona destinada a sumergir los sustratos que es de 21 mm de profundidad, está construida en policloruro de alta densidad e introducida en un armazón de aluminio para darle rigidez.

En la figura 1 se muestra un esquema general de la cuba: la barrera móvil, representada por B, es de teflón, se desplaza por encima de los bordes de la cubeta; está accionada por un motor de corriente continua MAXON R1543026 que puede ponerse en funcionamiento manualmente o a través de un ordenador utilizando un conector EUROCARD. La velocidad de desplazamiento de la barrera puede modificarse entre 0,02 y 1 mm s^{-1}.

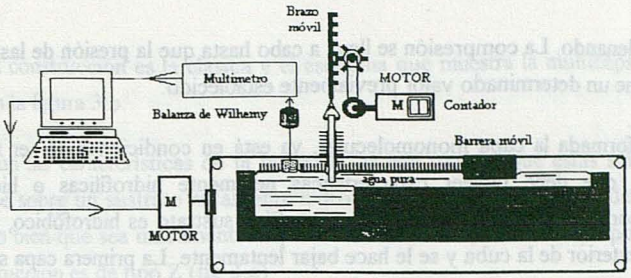


Figura 1 : Cuba de Langmuir

El brazo móvil o sistema de trasferencia de las capas monomoleculares sobre un sustrato está formado por una pinza de sujeción del sustrato, situada en el extremo de un brazo dotado de movimiento vertical y accionado por un motor, de las mismas características que el anterior, que permite una velocidad variable y una carrera de hasta 150 mm.

Un contador electrónico KCY-2 DM nos permite prefijar el número de capas a trasferir, y un conmutador ofrece la posibilidad de terminar con el sustrato introducido en la cubeta (si deseamos que nuestra superficie hidrófila), o bien en el aire, (cuando la superficie deseada deba ser hidrófoba).

Para la medida de la tensión superficial se utiliza como balanza de Wilhemy un transductor de desplazamiento continuo de marca SCHLUMBERG tipo PF.1.0 alimentado a 10 V por una fuente de corriente continua estabilizada. La sonda para la lectura de la presión superficial de las moléculas es una tira de papel especial de filtro de 20 mm x 20 mm.

El calibrado para la lectura de la presión se ha realizado con varias masas entre 10 y 300 mg, comprobando así la linealidad en la respuesta del transductor que se recoge en un multimetro Fluke modelo Hydra cuya precisión es de 10^{-4} V.

Técnica experimental de Langmuir-Blodgett.

En el proceso experimental hay que considerar dos etapas: la formación en la interfacie aire-agua de una capa monomolecular y la trasferencia de dicha capa sobre un sustrato sólido. La secuencia clásica para la realización de estas multicapas se muestra en la figura 2

a) Se realiza en primer lugar la dispersión de las moléculas sobre la superficie del agua, haciendo caer gota a gota una disolución muy diluida de la molécula objeto de estudio, quedebe poseer características anfifílicas, en un disolvente inmiscible con el agua, y volátil (v.g cloroformo). Una vez evaporado el disolvente la molécula permanece fijada sobre el agua ya que la cabeza polar la retiene en la superficie y la larga cadena alifática le impide introducirse en el interior del agua.

b) Situadas las moléculas en la superficie se procede a la formación de la capa, para ello hay que comprimir lateral, y muy lentamente la superficie de modo que las moléculas se vayan

aproximando y ordenando. La compresión se lleva a cabo hasta que la presión de las moléculas en la superficie tiene un determinado valor previamente establecido.

c) Una vez formada la capa monomolecular, ya está en condiciones de ser transferida sobre el sustrato, que debe poseer características netamente hidrofílicas o hidrofóbicas (generalmente se consigue con el método de lavado). Si el sustrato es hidrofóbico, se le sitúa inicialmente en el exterior de la cuba y se le hace bajar lentamente. La primera capa se forma al irse uniéndo la molécula al sustrato por la cadena, y al finalizar, el sustrato se halla introducido en el agua se encontrará recubierto por una capa de moléculas orientadas con la cabeza polar hacia el exterior.

d) Para efectuar la trasferencia de la siguiente capa, el sustrato debe elevarse lentamente, su superficie presenta en ese momento caracter hidrofilico y por lo tanto la molécula se adhiere por la cabeza polar quedando de nuevo recubierto con una nueva capa con las cadenas hidrofóbicas hacia el exterior. Siguiendo esta secuencia se va formando el edificio molecular.

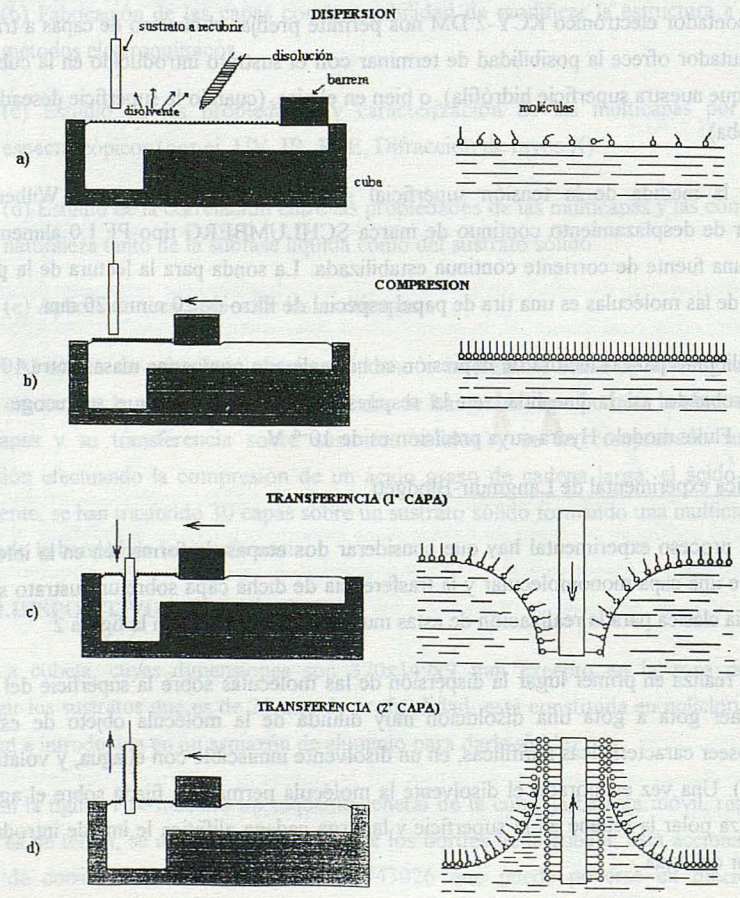


Figura 2: **Secuencia de la técnica LB.**

a) Dispersión de las moléculas b) Compresión lateral
c)Transferencia de la primera capa d) Transferencia de la segunda capa

Esta construcción es la clásica y el esquema que muestra la multicapa tipo Y, es el que aparece en la figura 3.b.

Según las características de la molécula, puede suceder que estas sólo sean capaces de transferirse sobre un sustrato netamente hidrófobo, y la construcción se denomina de tipo X, (fig 3.a), o bien que sea unicamente la parte hidrofilica la que se va superponiendo al sustrato, y la construcción es de tipo Z (fig 3.c).

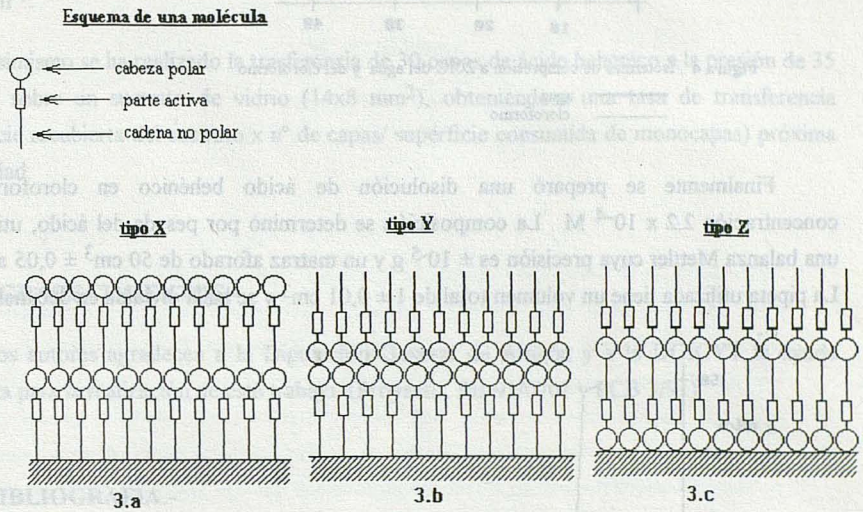


Figura 3: Diferentes tipos de apilamientos de las capas LB según el modo de transferenc

Comprobación del dispositivo experimental

Los productos utilizados para la obtención de la monocapa son:

Agua, calidad milli Q 18 MΩ de resistencia.

Cloroformo, para limpieza y disolución, Lab-Scan HPLC > 99'8%

Acido behénico (docosanoico) Fluka puriss > 99%

En primer lugar se comprobó la calidad del agua empleada así como el grado de limpieza de la superficie; para ello se efectuó un barrido desde la posición de 35 cm, que corresponde al extremo de la cubeta mas alejado de la balanza, hasta 3,5 cm del otro extremo. Se obtuvo una variación de la presión inferior a 1 $mN.m^{-1}$ aceptada generalmente como buena. La ausencia de impurezas en el cloroformo se contrastó mediante una operación de barrido similar, dispersando previamente sobre la superficie del agua 0,5 cm^3 de $CHCl_3$. Las compresiones obtenidas en ambos casos se muestran en la figura 4

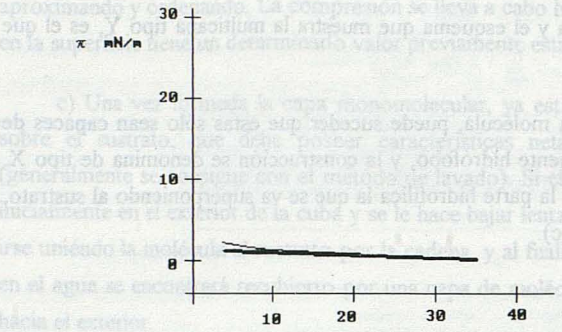


Figura 4 : Isotermas de compresión a 20°C del agua y del cloroformo

agua
cloroformo

Finalmemte se preparó una disolución de ácido behénico en cloroformo de concentración 2,2 x 10^{-4} M . La composición se determinó por pesada del ácido, utilizando una balanza Mettler cuya precisión es ± 10^{-5} g y un matraz aforado de 50 cm^3 ± 0,05 a 20° C. La pipeta utilizada tiene un volumen total de 1 ± 0,01 cm^3 y se halla dividida en décimas.

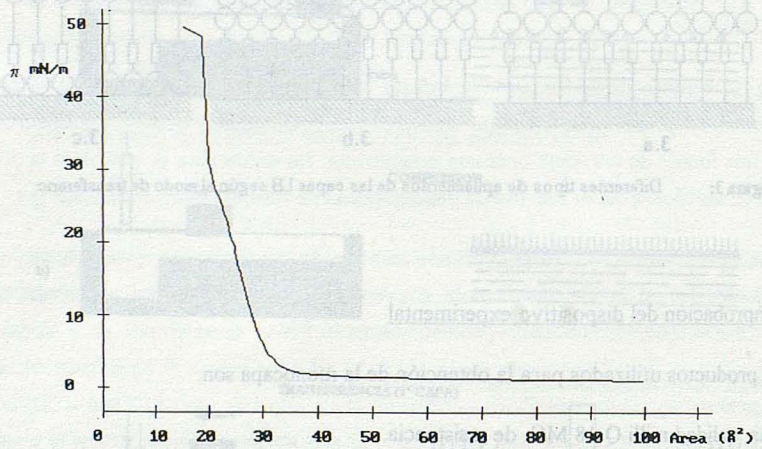


Figura 5 : Isoterma de compresión a 20°C del ácido behénico

La curva de compresión, presión vs área por molécula, fue reproducida tres veces en condiciones idénticas a 20±1 °C : 0,4 cm^3 de disolución dispersada gota a gota en la superficie de la cuba, esperando 10 minutos para la evaporación del disolvente.

La velocidad de avance de la barrera fue de 0,3 cm/min, es decir 0,8 $Å^2$/molécula/min.

En la fig 5 se representa la curva de compresión π-A. El área inicial para las moléculas está próxima a 100 $Å^2$, y a 30 $mN.m^{-1}$ su valor está próximo a 20 $Å^2$ que es el valor clásico que se obtiene para estas moléculas (6).

En condiciones similares (7) e incluyendo el error debido al experimentador, se estima que la imprecisión en el área de la molécula es de ±10% es decir ±2$Å^2$ $molec^{-1}$

El colapso de la superficie, en estas condiciones tiene lugar a presiones próximas a 50 $mN.m^{-1}$.

Asimismo se ha realizado la trasferencia de 30 capas de ácido behénico a la presión de 35 $mN.m^{-1}$ sobre un sustrato de vidrio (14x8 mm^2), obteniendose una tasa de transferencia (superficie recubierta del sustrato x nº de capas/ superficie consumida de monocapas) próxima a la unidad.

AGRADECIMIENTOS.-

Los autores agradecen a la Diputación General de Aragón y a la DGICYT la ayuda facilitada para la realización de este trabajo. (Proyecto, PB 91/0701 y PCB 3/91)

BIBLIOGRAFIA.-

1.- Vandevyver, M.; Barraud, A.; Lesieur, P.; Richard, J.; y Ruaudel-Teixier, A.; Journal de Chimie Physique, 83, 1986 (1986)

2.- Tieke, B.; Wegmann, A.; Fischer, W.; Hilti, B.; Mayer, C.W.; y Pfeiffer, J.; Thin Solid Films, 179, 233 (1989)

3.- Langmuir, I.; Trans. Faraday Society, 15, 62 (1920)

4.- Blodgett, K.; J.Am. Chem. Soc. 57, 1007 (1935)

5.- Kuhn, K.; Z. Naturferch 17, A 411 (1962)

6.- Adamson, A.; "Physical Chemistry of Surfaces" . John Wiley & Sons, Inc. N.Y. (5ª ed.) Cap IV. (1990)

7.- Céline Dourthe ; Tesis de Doctorado E.N.S.C.P.B. 1991 Univ. de Bordeaux

Rev. Academia de Ciencias. Zaragoza **48** *(1993).*

SEGUIMIENTO FOTOGRAFICO DE LA EROSION EN LA VAL DE LAS LENAS (ZARAGOZA). ESTUDIO PRELIMINAR.

M.A. Soriano

Departamento de Geología.
Universidad de Zaragoza. 50009 Zaragoza. España

Abstract

The Ebro basin was filled by detrital, gypsum and carbonate deposits of a Tertiary age. During the Quaternary different landforms were developed; the more recent are the infilled valleys, which are all around the central Ebro basin. Accumulative and erosion periods have cooperated for its development. A periodic control by means of photographs carried out for four years shows that there are small changes in the bottom of the valleys because of the gully activity. In the slopes where piping is acting, the variations observed are important. Collapses caused by piping and fluvial erosion are also actives processes. From our experience it seems to be that monitoring infilled valleys using photographs is a useful tool for knowing the evolution of these landforms.

1. Introducción.

Los valles de fondo plano o "vales" son uno de los modelados más abundantes que se encuentran en el centro de la Depresión del Ebro. Han sido rellenados parcialmente por materiales detríticos, estando en su mayoría erosionados por incisión fluvial y procesos de piping. Es frecuente que en estos valles se hayan producido varias etapas de sedimentación separadas por sendos episodios erosivos. Los estudios que se han llevado a cabo sobre los mismos no son muy numerosos y se centran bien en determinar su génesis (Llamas, 1962; Torras y Riba, 1968; Soriano y Calvo, 1987; Cuchí y Soriano, 1993), bien en establecer cuántos y qué edad tienen los niveles acumulativos que se han producido comparándolo con lo que ocurre en el resto del área mediterránea (van Zuidam, 1975; Burillo et al., 1985; 1986; Cuchí y Soriano, 1993).

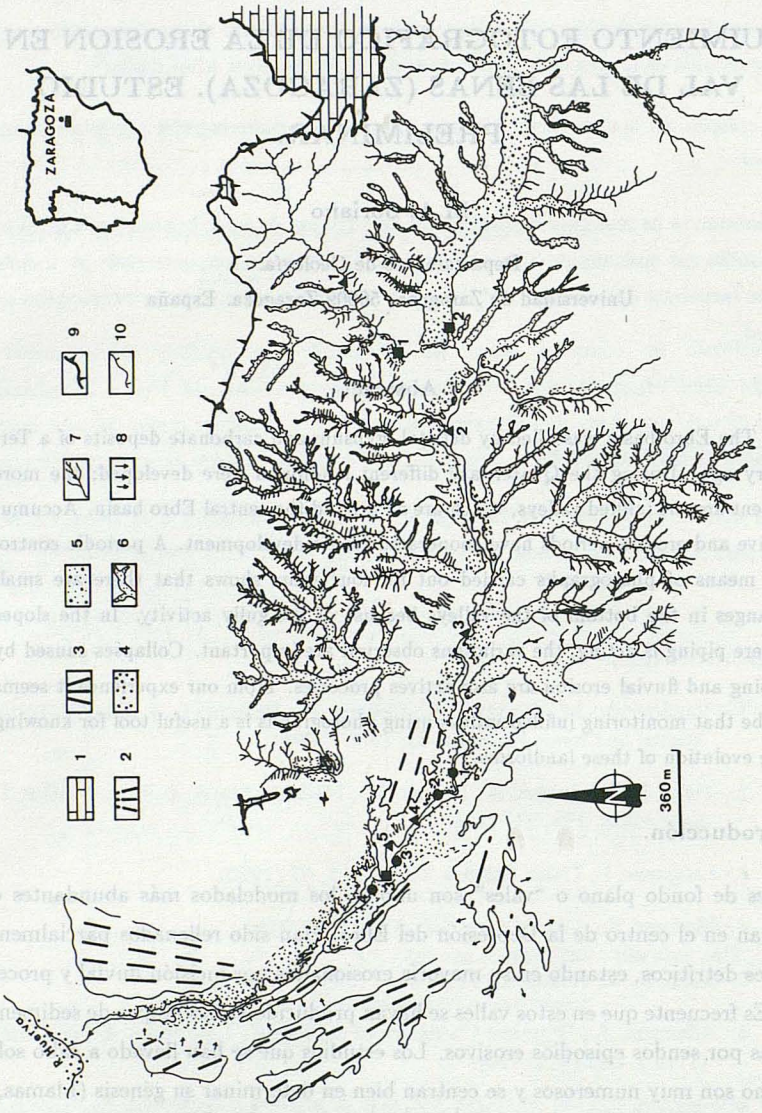


Figura 1.- Esquema geomorfológico y de situación de la val de las Lenas. 1. Relieve estructural en calizas terciarias. 2. Glacis 6. 3. Glacis 3. 4. Terraza 1. 5. Relleno de la val de las Lenas. 6. Valles de fondo plano. 7. Incisión lineal. 8. Regulación de vertiente. 9. Escarpe estructural en Terciario. 10. Escarpe en Cuaternario. A lo largo del valle se ha situado la localización de los perfiles realizados. con los siguientes símbolos: ▲ Perfiles transversales del barranco, ● Piping y ■ Colapsos.

En este trabajo se ha estudiado con más detalle la val de las Lenas que se localiza en las proximidades de las poblaciones de Botorrita y de María de Huerva (figura 1). Esta zona se encuentra en el centro de la Depresión del Ebro donde los materiales aflorantes son de edad neógena y están constituidos por facies detríticas, evaporíticas y carbonatadas. Durante el Cuaternario la red fluvial ocasionó, en los valles principales, la formación de ocho niveles encajados de terrazas y glacis. Durante el Holoceno la alternancia de periodos en que domina la sedimentación y la erosión produjo el desarrollo de hasta cuatro niveles acumulativos en los valles de fondo plano (Cuchí y Soriano, 1993) y de depósitos de vertiente. Las acumulaciones de los valles están relacionados con el nivel de terraza más reciente de los cauces fluviales de la zona (Ebro y Huerva, principalmente).

En este trabajo se han realizado fotografías periódicas de varios perfiles localizados a lo largo de un valle de fondo plano con el objetivo de determinar si se producen cambios en los mismos. Si este método se revela útil para ello, se podran utilizar otros más complejos para evaluar y cuantificar dichas variaciones. Además, también se expondrán las principales características morfológicas, sedimentológicas y cronológicas de estos valles.

2. Características fundamentales.

Los valles de fondo plano disectan los materiales terciarios de esta región y su trazado es meandriforme formando redes dendríticas. La erosión actual a que se hallan sometidos permite, en la mayoría de los casos, determinar las características sedimentológicas de los rellenos. En buena parte de ellos se aprecian distintos niveles encajados entre sí. En algún caso se han encontrado cuatro acumulaciones e incluso una quinta dudosa (Cuchí y Soriano, 1993), pero en la mayoría de ellos se encuentran un máximo de tres niveles, cuya altura del más antiguo al más moderno sobre el actual cauce del barranco es de 10-14, 3-4 y 1-2 m, respectivamente (Soriano y Calvo, 1987 y Soriano, 1989).

Por lo general, los sedimentos presentan grandes variaciones en su granulometría y estructuras sedimentarias tanto en vertical como en horizontal. Están integrados por limos y gravas, siendo más abundanes los primeros. Entre las estructuras sedimentarias se pueden señalar: laminación horizontal, cruzada, ripples, cantos imbricados, canales, laminación flaser, convoluta, grietas de desecación, marcas de erosión, tool marks, costras salinas, gotas de lluvia, huellas de animales, superficies erosivas sobre Terciario y dentro del relleno holoceno, etc. Las primeras son comunes en los niveles antiguos y en los recientes, mientras que las marcas superficiales sólo se han observado en el más moderno.

Del tipo de estructuras encontradas en el nivel actual se deduce que el origen de las mismas serían corrientes efímeras (Picard y High, 1973). El hecho de que algunas de ellas sean comunes en los distintos niveles indicará que los procesos que intervinieron en

la sedimentación de ellos son similares a los observados en la actualidad para el nivel inferior. Otro aspecto a tener en cuenta es la presencia de superficies erosivas dentro del relleno cuaternario sobre las que hay, fundamentalmente, limos de vertiente, con lo que resulta evidente que también existen aportes de materiales de esta procedencia a la formación de estos rellenos si bien en menor medida que los de origen fluvial.

La edad de estas acumulaciones se puede determinar de forma relativa y de forma absoluta. Así, Soriano (1989) a partir de restos arqueológicos existentes indica que la parte superior del depósito más antiguo en la zona de María de Huerva-Monasterio de Santa Fé debe ser postromana-previsigoda. A partir de C14 Cuchí y Soriano (1993) señalan que la edad de la parte superior del relleno que encuentra a 6 m sobre el cauce actual del barranco en una val próxima a Torrecilla de Valmadrid es de hace 2000±80 y 2140±220 años. Por otra parte, en el valle del Huerva, en el relleno superior de una val próxima al Monasterio de Santa Fé se ha obtenido una nueva datación absoluta de la parte superior del depósito que indica que se formó hace 3750±80 años.

3. Val de las Lenas.

Para realizar este estudio se ha elegido la val de las Lenas que se sitúa entre las localidades de María de Huerva y de Botorrita. Este es uno de los valles de fondo plano con mayor longitud de esta zona y presenta buena parte de su trazado incidido por un barranco (figura 1). Por estos motivos se pensó que sería un buen valle para observar posibles cambios como consecuencia de la acción de procesos erosivos que actúan sobre los sedimentos del relleno y en el fondo del barranco actual. Además a lo largo del perfil longitudinal del valle se pueden producir variaciones en la relación erosión-sedimentación.

El seguimiento fotográfico que se ha efectuado ha sido periódico y sistemático a lo largo de cuatro años, realizándose trimestralmente desde la primavera de 1989 hasta la de 1993. A partir de las diapositivas tomadas se elaboran sendos esquemas para facilitar la comparación de las imágenes sucesivas. Para ello, es preciso que el lugar y el ángulo con el que se toma la imagen no cambie, si bien nuestra experiencia personal muestra que en muchas ocasiones es difícil poder cumplir con ello ya que las marcas puestas para señalar los distintos puntos han sido arrancadas. Además como apoyo, se han utilizado registros fotográficos más puntuales y esporádicos que abarcan desde 1984 hasta 1989.

Las imágenes tomadas se han centrado en analizar posibles cambios en tres aspectos principales: (1) perfil transversal del barranco (2) estructuras causadas por piping y (3) estructuras de colapsos. Se ha centrado la toma de imágenes en estos aspectos ya que son numerosos los autores que señalan que todos o parte de estos procesos erosivos son los más importantes que intervienen en el desarrollo del abarrancamiento (Dardis, 1989; López-

Bermúdez y Romero-Díaz, 1989 y Oostwoud y Bryan, 1991). En la figura 1 se encuentran señalados todos los lugares elegidos a lo largo del valle de las Lenas y el carácter dominante que se ha observado en cada uno de ellos.

3.1 Perfil transversal del barranco.

Se han analizado cinco perfiles que se encuentran distribuidos de forma bastante regular a lo largo del cauce del barranco actual (ver figura 1). Tan solo uno de ellos se sitúa en un valle lateral próximo a donde se inicia la erosión lineal de la val y donde no hay relleno en el valle; los demás se localizan en el barranco que incide a la val.

En general, los cambios observados en las paredes del valle son muy escasos. En la zona de cabecera donde el barranco se encaja sobre sedimentos yesíferos, no se ha apreciado ninguna variación (perfil n. 1, figura 1). Aguas abajo, donde el barranco se encaja bien sobre Terciario detrítico, bien sobre los sedimentos cuaternarios de relleno, se han visto variaciones tenues en alguna de las pequeñas concavidades que se encuentran en las vertientes. En el fondo del valle queda registrada buena parte de la actividad erosivo-sedimentaria que se produce a lo largo del mismo. Sin embargo, las nuevas incisiones y acumulaciones que representan el signo evidente de dicha evolución son, en muchos casos, difíciles de comprobar debido a la presencia de vegetación estacional que se instala aprovechando las zonas de mayor humedad y que cubre a los sedimentos. A pesar de ello, en varios de los perfiles (3 y 5, figura 1) se aprecia el crecimiento de barras y también de pequeñas incisiones en los últimos años que no se habían visto en los primeros. Hay que hacer notar que el crecimento de la barra del perfil 3 parece estar relacionado con una modificación llevada a cabo por actividad humana en los caminos que cruzan dicho barranco a unos 200 m aguas arriba de donde se ha venido realizando el control.

3.2 Piping.

Este proceso se produce en toda la zona donde hay relleno cuaternario de la val, si bien es en la parte final de la misma donde su desarrollo es más espectacular y, por lo tanto, donde se han centrado especialmente los estudios (ver figura 1). Las morfologías observadas a causa del piping en esta zona son dispares: conductos verticales, horizontales, pequeños hundimientos, pseudodolinas, puentes, estructuras turriculadas, etc. A pesar de que estos modelados se generaron inicialmente por dicho proceso, en la actualidad intervienen en su evolución otros procesos, tales como colapsos y erosión fluvial, como se verá a continuación. Como ejemplo de las variaciones producidas a lo largo de estos años se muestra el puente de la figura 2.a y 2.b (punto n. 3, figura 1). En él se aprecia claramente un aumento de sus dimensiones. De esta manera la altura del techo en su zona central se ha visto

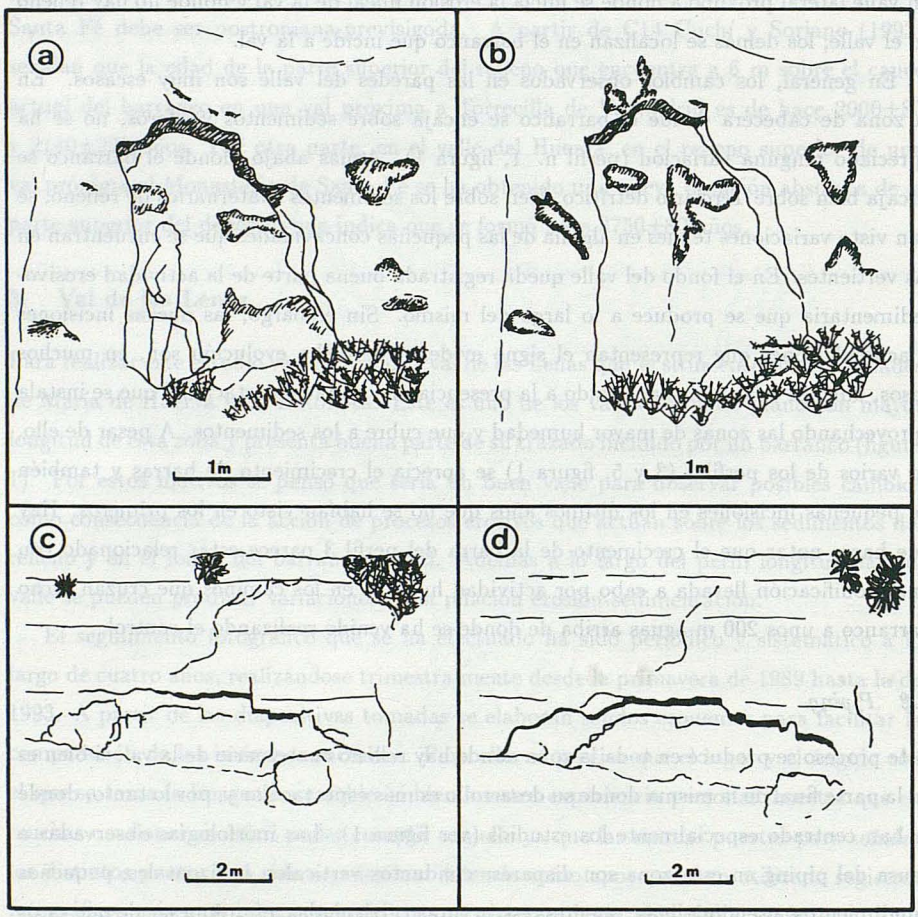


Figura 2.- Esquemas mostrando las variaciones sufridas por alguno de los modelados de la val de las Lenas. (a) y (b) Cambios en un puente formado por piping representado por el punto 3 de la figura 1, (a) Noviembre de 1988, (b) Septiembre de 1992. (c) y (d) Modificaciones del colapso señalado con el punto 3 en la figura 1 y que se muestra en la figura 3.c y 3.d, (c) Abril de 1989, (d) Mayo de 1993.

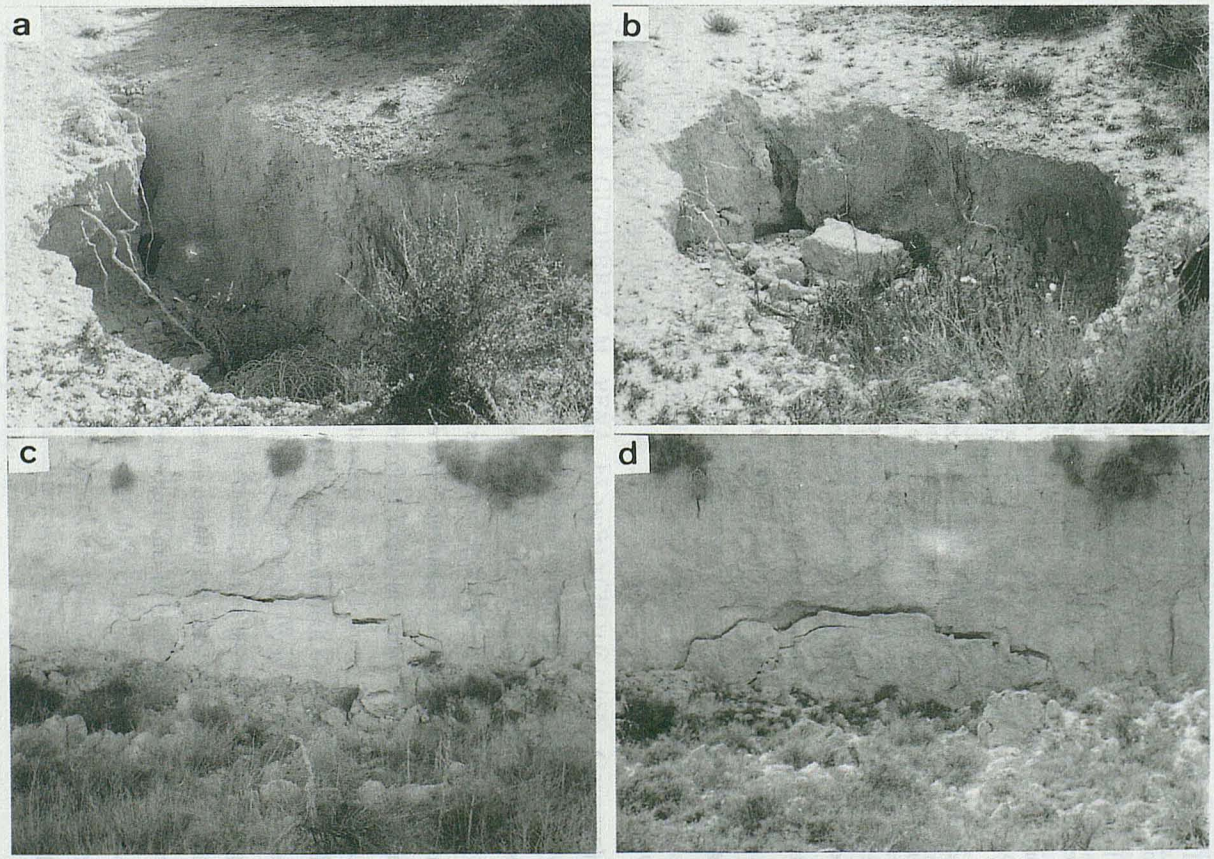


Figura 3.- Variaciones en los modelados del valle de las Lenas. (a) y (b) Cambios en una depresión generada por piping,, punto 4 de la figura 1. (a) Abril de 1992, (b) Mayo de 1993 . (c) y (d) Modificaciones en el colapso señalado en el punto 3 de la figura 1, (c) Marzo de 1989, (d) Mayo de 1993.

incrementada en aproximadamente 14 cm en un periodo de cuatro años (desde 1988 a 1992), lo que indica que el promedio es de 3,5 cm/año. Ello es debido, sobre todo, a causa de desprendimientos de bloques de las paredes. El material que se encontraba caído en el suelo en el momento de efectuar una de las primeras observaciones ha desaparecido en la actualidad. Las zonas hundidas que podían verse en las paredes del material de relleno de la val y que eran restos de otros conductos activos anteriormente, han ido suavizando esa forma cóncava debido a la erosión continuada de las mismas. A finales de 1992 se detectó la presencia de una nueva cavidad abierta en la base de este mismo puente como consecuencia de la evolución de un conducto interno.

A comienzos del año 1991 se produjeron unos colapsos de hasta un metro de diámetro máximo y unos ochenta centímetros de profundidad, causados por el hundimiento de conductos en una de las sendas por las que se accede al fondo del valle (punto n. 4, figura 1). lo que demuestra la actividad continua de este proceso. En las imágenes mostradas en la figura 3.a y 3.b se aprecia claramente el incremento de los diámetros sufrido por una de las depresiones en un año (de abril de 1992 a mayo de 1993), si bien en 1993 (figura 3.b), su profundidad era menor a causa de la acumulación en el fondo de material caido de las paredes. Este hecho produce que no se distingua con tanta nitidez el conducto interno que la originó (parte izquierda de la zona hundida).

3.3 Colapsos.

Se originan cuando parte del material infrayacente desaparece y, consecuentemente, se produce una caida de los sedimentos superiores. Esta falta de sustentación se produce por erosión (fluvial o por piping) de los niveles, fundamentalmente, cuaternarios que constituyen el relleno alto de la val de las Lenas. Como ocurre con el piping, donde se observan mejor los colapsos es en aquellos lugares donde los materiales de relleno de la val son más potentes (ver figura 1) y, por tanto, de forma especial en la zona terminal de la misma (punto n. 3 en figura 1). Por lo general se han apreciado cambios importantes en aquellos colapsos que se han observado. De esta forma en el ejemplo representado en los esquemas y la fotografía de las figuras 2.c, 2.d, 3.c y 3.d se aprecia claramente cómo algunas grietas que se veían incipientes en 1989 tienen la misma altura que la grieta principal que desplaza al núcleo del colapso en 1993 (entre 25 y 28 cm de altura), con lo el incremento de ésta es de unos 14 cm. Por término medio el aumento de la anchura de las grietas esta comprendido entre 5 y 10 cm en este periodo de tiempo (lo que representa un promedio del1,25 a 2,5 cm/año). No se ha producido una aumento apreciable en la longitud de las grietas.

4. Conclusiones.

Los valles de fondo plano constituyen uno de los modelados más frecuentes encontrados en el centro de la Depresión del Ebro. Mediante el control fotográfico de un valle de fondo plano (val de las Lenas) a lo largo de cuatro años, ha sido posible detectar en varios de los perfiles y zonas analizadas variaciones, en ocasiones importantes, que revelan la utilidad de esta técnica para establecer una primera aproximación acerca de zonas en las que los procesos erosivos sean activos y en las que posteriormente se pueda aplicar otro tipo de seguimiento más sofisticado. Además, en el valle estudiado se observa que si bien los procesos fluviales tienen cierta actividad, los que producen las variaciones más importantes (especialmente en las zonas en que existe un relleno de valle antiguo) son claramente los de piping, observándose durante este tiempo el desarrollo de formas nuevas que no se habían detectado cuando se comenzó a llevar a cabo el seguimiento fotográfico en 1989. El colapso de materiales de las vertientes de los valles también presenta variaciones facilmente detectables en este periodo de tiempo.

Referencias

BURILLO, F., GUTIERREZ, M. y PEÑA, J.L. (1985) Las acumulaciones holocenas y su datación arqueológica en Mediana de Aragón (Zaragoza).*Cuadernos de investigación Geográfica*, t. **XI**, pp. 193-207.

BURILLO, F; GUTIERREZ, M., PEÑA, J.L. y SANCHO, C.(1986). Geomorphological processes as indicators of climatic changes during the Holocene in the North-East Spain. *Quaternary climate in western mediterranean*, pp. 31-4.

CUCHI, J.A. y SORIANO. M.A. (1993) The val of San Marcos. Study of an infilled valley in the Ebro basin (Spain). *Zeitscrift fur Geomorphology* (en prensa).

DARDIS, G.F. (1989) Quaternary erosion and sedimentation in badland areas of southern Africa. *Catena Supp.* **14**, pp. 1-19.

LOPEZ-BERMUDEZ, F. y ROMERO-DIAZ, M.A. (1989) Piping erosion and badland development in South-east Spain. *Catena Supp.* **14**, pp. 59-73.

LLAMAS, R.M. (1962) Estudio geológico-técnico de los terrenos yesíferos de la cuenca del Ebro y de los problemas que palantean en los canales Servicio Geológico.*Ministerio de Obras Públicas informaciones y estudios*, bol n.**12.** 192 p. Madrid.

OOSTWOUD WIJDENES, D.J. y BRYAN, R.B. (1991) Gully development on the Njemps Flats, Baringo, Kenya. *Catena Supp.* **19**, pp. 71-90.

PICARD, F. y HIGH, L.R. (1973) Sedimentary structures of ephemeral streams. *Developments in Sedimentology*, **17**, 223p. Elsevier.

TORRAS, A. y RIBA, O. (1968) Contribución al estudio de los limos yesíferos del centro de la Depresión del Ebro.*Boletín del instituto de Estudios Asturianos*, **14**.

SORIANO, M.A. y CALVO, J.M. (1987). Características, correlación y datación de los valles de fondo plano de Zaragoza. *Geomorfología y Cuaternario* **1**, pp. 283-293.

SORIANO, M.A. (1989) Infilled valleys in the central Ebro basin (Spain). *Catena* **16**, pp. 357-367.

ZUIDAM van, R.A. (1975) Geomorphology and Archaeology. Evidences of interrelation at historical sites in the Zaragoza region, Spain. *Zeitscrift fur Geomorphology*, **19**, pp. 319-328.

Rev. Academia de Ciencias. Zaragoza. 48 (1993)

CATODOLUMINISCENCIA DE LAS DOLOMÍAS DE LA FORMACIÓN RIBOTA (CADENA IBÉRICA ORIENTAL, ESPAÑA).

A.J. Zamora, J. Mandado, J.M. Tena, L.F. Auqué y M.J. Gimeno

Departamento de Geología. Facultad de Ciencias. Universidad de Zaragoza. 50009 Zaragoza.

ABSTRACT

Lower Cambrian carbonate materials of Ribota Formation show important diagenetic processes superimposed over their primary features. Such processes prevent or make difficult the use of conventional petrographic techniques to study those carbonate materials.

The use of cathodoluminiscence (CL) technic as petrographic tool on these materials allows to evaluate compositional variations of minor elements (mainly Fe and Mn) complementing other techniques (transmitted and polarized light petrography and/or geochemical trend analysis) and furnishing a posterior genetic interpretation.

Study by means of CL techniques allows to recognize sinsedimentary relict structures (not observed with conventional petrographic ones) and the homogeneity of dolomitization processes in those materials. Other diagenetic processes, such as recrystallization, epidiagenetic fracturation-cementation phenomena and dedolomitization can be recognized by CL techniques too. Cathodoluminiscence analysis consolidates the petrographic model proposed for these carbonate materials.

1. INTRODUCCIÓN

El estudio petrológico de los materiales carbonatados de la Formación Ribota presenta gran cantidad de dificultades para determinar tanto la composición y tipo de los sedimentos originales, como para establecer un modelo de las transformaciones postsedimentarias sufridas por los mismos. La superposición de los diferentes procesos diagenéticos, entre los que hay que destacar como más importantes los de dolomitización y recristalización (Zamora *et al.*, 1992), genera unas litologías muy uniformes, caracterizadas por la presencia de mosaicos cristalinos esparíticos en los que es sumamente difícil reconocer componentes texturales primarios.

El uso de la catodoluminiscencia (expresada en general de forma abreviada como CL por razones de comodidad) permite solventar algunas de estas lagunas y complementa la información adquirida mediante las técnicas petrográficas convencionales.

La catodoluminiscencia se basa en el fenómeno de emisión de luz de un material, en nuestro caso superficie pulida de una muestra de roca carbonatada, cuando se hace incidir sobre ella un haz de electrones energéticos.

El fenómeno físico consiste básicamente en que los electrones incidentes al chocar con los electrones de los átomos, les confieren parte de su energía cinética, promocionándolos a un orbital o nivel de energía "potencial" más elevado y trans-

formándolos en átomos excitados. Posteriormente, el electrón al volver a su nivel energético normal emite la energía absorbida en forma de radiación luminosa y calor.

La luz que emite cada mineral tiene un determinado valor máximo de longitud de onda; es decir, tiene una luminiscencia máxima para un valor o intervalos de valores de longitud de onda, que son específicos de cada mineral.

Hay dos conceptos o términos que se asocian al caracterizar la luminiscencia y que son: el color (considerando sólo las radiaciones emitidas en el espectro visible) y el grado de luminiscencia o intensidad de la luminiscencia.

En los carbonatos la razón de la luminiscencia o no luminiscencia de los mismos no depende sólo de la naturaleza mineralógica de los cristales (luminiscencia intrínseca), sino también del tipo de elementos traza presentes en los mismos y de la posición específica que ocupan en la red (luminiscencia extrínseca). Los elementos más importantes desde este punto de vista son el manganeso, como elemento potenciador de la luminiscencia, y el hierro, como inhibidor de la misma. Ambos elementos se encuentran en las redes de los carbonatos sustituyendo a los cationes principales. En las dolomías el Fe y Mn ocupan preferentemente las posiciones del magnesio debido a la mayor similitud de radios atómicos (Pierson, 1981).

Independientemente de las variaciones de intensidad de la luminiscencia, la identificación de los distintos carbonatos en CL se basa en los diferentes espectros de color de cada mineral, aunque hay importantes variaciones de color en las diferentes especies mineralógicas carbonatadas como consecuencia de la influencia de los factores geoquímicos y estructurales. Así, la calcita presenta por lo general una luminiscencia intensa naranja y la dolomita menos intensa y rojiza.

2. METODOLOGÍA

La técnica de la CL se ha aplicado sobre las superficies pulidas de muestras de rocas carbonatadas de la Fm. Ribota, utilizando un cañón de electrones de la marca THECHNOSYN, modelo 8200 MK II, acoplado a un microscopio NIKON y bajo las siguientes condiciones de trabajo: la intensidad de corriente para obtener resultados positivos se fijó en 150-200 μA y con una diferencia de potencial de 10-13 Kv.

Hay que resaltar un hecho que se produce en las muestras estudiadas y que consiste en que la mayoría de ellas no son luminiscentes en las condiciones de trabajo habituales descritas en libros y artículos de revistas (10-12 Kv y 500 μA); sin embargo, para condiciones de 10-13 Kv y 150-200 μA sí que presentan luminiscencia, observándose además que al aumentar la intensidad de la corriente del chorro de electrones dicha luminiscencia va desapareciendo. Este hecho precisa de un estudio más detallado para poder darle una explicación, aunque es preciso mencionar que algunos autores citan condiciones de trabajo muy similares a las nuestras (como es el caso de Marshall, 1988, que indica condiciones de trabajo de 12 Kv y 140 μA).

3. CONSIDERACIONES GEOQUÍMICAS SOBRE LA LUMINISCENCIA DE LOS CARBONATOS DE LA FM. RIBOTA.

Las muestras analizadas, al estudiarlas mediante CL, se pueden agrupar en tres conjuntos no muy bien diferenciados y de límites difusos entre ellos, en función de la presencia o ausencia de luminiscencia y de la intensidad de la misma. En el conjunto de las muestras podemos distinguir algunos carbonatos sin luminiscencia

aparente (luminiscencia negra o no luminiscentes), otros con luminiscencia tenue a muy tenue y, finalmente, aquellos de luminiscencia media a intensa.

El color de la luminiscencia en todas ellas es rojo profundo (principalmente) con variaciones a rojo anaranjado, lo que indica que la composición del carbonato (sea micrítico, microsparítico o esparítico) sería muy próxima a la composición de la dolomita (Pierson, 1977).

Basándonos en esta agrupación de las muestras y en los datos analíticos de las mismas (tabla 1), hemos intentado reflejar la relación entre las características de la luminiscencia y la composición química de las distintas muestras. Para ello se ha realizado un diagrama XY (figura 1) utilizando los datos del contenido en Fe y Mn (expresados en tanto por ciento), que son respectivamente el elemento inhibidor y potenciador del tipo e intensidad de la luminiscencia en los carbonatos según la mayoría de los investigadores (Oglesby, 1976; Pierson, 1981; Amieux, 1982; Fairchild, 1983; y Have y Heijnen, 1985, entre otros muchos).

De la representación de las distintos tipos cualitativos de luminiscencia observados, en base a los porcentajes de los cationes ya citados, se pueden alcanzar algunas conclusiones previas sobre la composición mineralógica de estas rocas y la influencia de los elementos traza más significativos.

No hay una separación clara entre dolomías no luminiscentes y de luminiscencia tenue, representándose todas las muestras de esas dos clases en una banda dentro de la nube de puntos. Por el contrario, sí que se puede establecer una neta diferenciación entre las dolomías con luminiscencia intensa y el resto de dolomías (no luminiscentes y luminiscencia tenue), a excepción de dos muestras que se proyectan en esa banda y que tienen una luminiscencia intensa, hecho éste que puede explicarse si tenemos en cuenta el hecho de que se trata de dos muestras de calizas puras y por tanto no son comparables en su comportamiento con las dolomías. Esta separación queda también manifiesta en los valores medios obtenidos para el Fe en las distintas clases (ver tabla 1).

TABLA 1. Valores medios (χ_m) y desviación estándar (σ_x) de los parámetros químicos analizados, para los tres tipos de luminiscencia diferenciados.

	No luminiscentes		**Luminiscencia tenue**		**Lum. media-intensa**	
	χ_m	σ_x	χ_m	σ_x	χ_m	σ_x
% CO_3	65.526	4.605	65.853	5.233	62.944	**2.681**
% Ca	19.366	5.592	19.814	5.699	24.576	6.328
% Mg	12.626	2.374	12.632	1.469	10.931	4.275
% Sr	0.003	0.004	0.008	0.009	0.011	0.011
% Fe	2.148	0.710	1.465	0.697	1.295	1.409
% Mn	0.213	0.108	0.135	0.100	0.139	0.093
% Na	0.064	0.333	0.065	0.049	0.077	0.045
% K	0.051	0.073	0.026	0.022	0.024	0.040

Las muestras no luminiscentes y de luminiscencia tenue presentan los contenidos más altos de magnesio, lo que indica que corresponden a los términos más dolomíticos, y por lo tanto menos luminiscentes. El contenido en Fe justifica también la respuesta a la CL; este elemento, inhibidor de la luminiscencia, muestra un claro incremento desde las más luminiscentes a las no luminiscentes, lo que concuerda con los supuestos teóricos. Por el contrario, el Mn, que es el elemento activador de la luminiscencia en los carbonatos, presenta sus máximos contenidos en las no luminiscentes, discrepancia que podría justificarse si tenemos en cuenta que en ellas el contenido en hierro es considerablemente más elevado que en el resto de las muestras.

El análisis detallado de la figura 1 permite justificar los supuestos establecidos anteriormente. La banda de valores de la relación Mn/Fe para las muestras no luminiscentes y de luminiscencia tenue, se ajusta a una función exponencial del tipo siguiente:

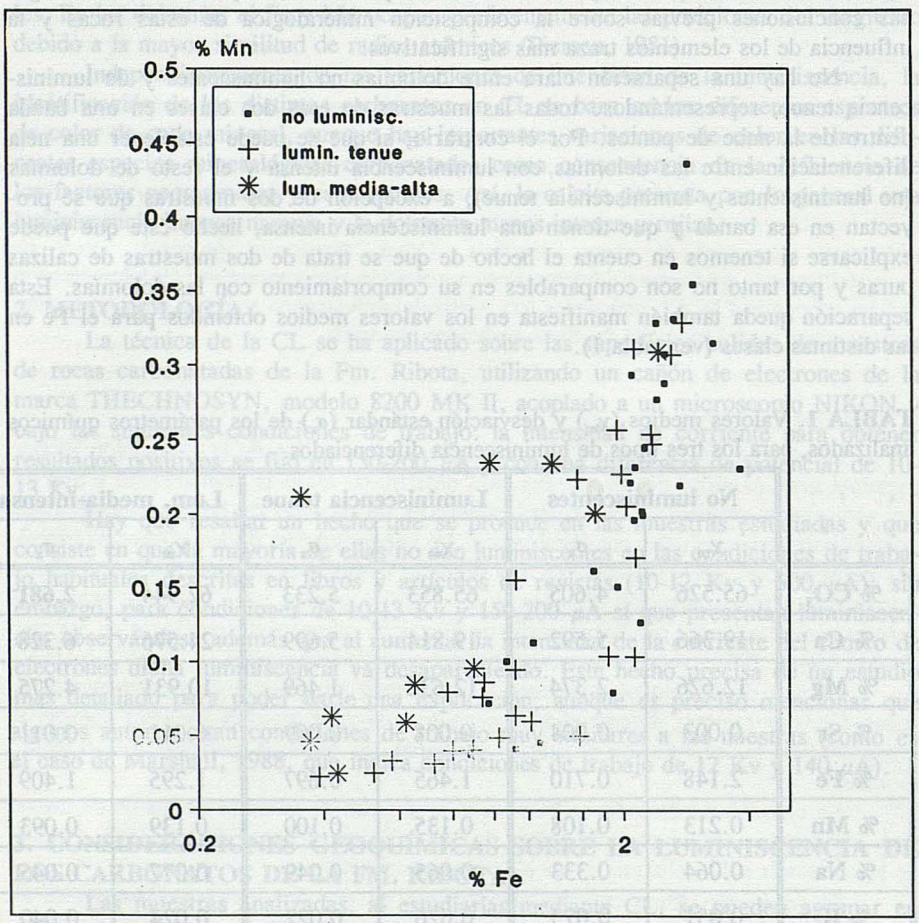


Figura 1. Relación de la luminiscencia de los carbonatos de la Fm. Ribota con el contenido de hierro y manganeso.

$$Mn = \exp [-2.43(\pm 0.071) + 2.49(\pm 0.246) * \log Fe]$$

y con un coeficiente de correlación de 0.7894. La parte izquierda de esta función para valores comprendidos entre 1.5 y 2 % de Fe es muy similar al límite inferior que estableció Fairchild (1983) para las dolomías precámbricas de la Formación Bonahaven de Escocia, con la salvedad de que este autor considera el límite como una línea recta con pendiente cero que corta al eje de ordenadas entre 0.015 y 0.03 % de Mn. Por encima del 2% de Fe, la función se asemeja mucho al límite establecido por Pierson (1981) para dolomías desde Cámbricas a Cretácicas (aunque este autor considera vertical), o al definido por Fairchild (1983).

De todo esto se deduce que, en las dolomías de la Formación Ribota, la luminiscencia es función de la relación Mn/Fe; teniendo más importancia el contenido en manganeso para valores de hierro inferiores al 2%, mientras que por encima de esos valores pasa a ser el contenido en hierro el factor casi primordial.

Para confirmar la validez de este planteamiento se ha realizado el análisis discriminante de los datos geoquímicos, utilizando como variables los contenidos de CO_3, Ca, Mg, Sr, Fe, Mn, Na y K. En un segundo caso procedimos a eliminar los datos de CO_3, Na y K, utilizando exclusivamente el resto de variables. En ambos casos se encuentran dos funciones que explican la totalidad de la variabilidad de la luminiscencia en nuestras muestras (tabla 2).

TABLA 2. Análisis discriminante. Caso 1: utilizando los contenidos de CO_3, Ca, Mg, Sr, Fe, Mn, Na y K. Caso 2: utilizando únicamente los de Ca, Mg, Sr, Fe y Mn. En negrita parámetros químicos más importantes de cada función discriminante.

	CASO 1		CASO 2	
	Func. Discr. 1	**Func. Discr. 2**	**Func. Discr. 1**	**Func. Discr. 2**
% CO_3	**-50.3924**	**30.4219**		
% Ca	**-62.8683**	**38.6459**	**-0.4276**	**0.5301**
% Mg	**-27.7833**	**16.2824**	0.1769	-0.4219
% Sr	0.6791	-0.4142	-0.3506	-0.3302
% Fe	**-10.0410**	**5.0930**	**0.7522**	-0.2272
% Mn	**-1.0707**	**1.2609**	0.0540	**0.5283**
% Na	-0.1653	0.4093		
% K	-0.7981	0.7954		
% Variación explicada	86.2	13.8	86.85	13.15
Eigenvalue	0.5071	0.0812	0.4192	0.0634

Para el CASO 1, la primera función discriminante explica el 86,2 % de la variación y en ella las variables de más peso son los elementos mayoritarios (CO_3, Ca y Mg) y el Fe y Mn; la segunda lo hace del 13,8 % y al igual que la anterior las

variables de más peso son los elementos mayores (CO_3, Ca y Mg) y el Fe y Mn. En este supuesto, la influencia de los contenidos de CO_3, Ca y Mg, elementos mayoritarios e interrelacionados en la composición del carbonato, oscurece la de los demás elementos. Para el CASO 2, en el que se eliminan los dátos del contenido en CO_3, ya relacionado con los del Ca y Mg, y de los alcalinos Na y K, la discriminación es algo mejor, obteniéndose una función discriminante que explica 86.85 % de la variación, en la que la variable de más peso es el Fe, y en menor proporción el Ca, y una segunda función discriminante que explica el 13,15 % de las muestras, siendo el Mn y el Ca las variables de más peso, pudiéndose interpretar que la primera agrupación incluiría las muestras con luminiscencia nula a tenue y la segunda a las de luminiscencia media a alta, dadas las variables de mayor peso en cada una de ellas. En ambos casos se comprueba cómo el Fe y el Mn son los que condicionan la luminiscencia extrínseca y que el Fe tendría un mayor peso que el Mn a la hora de explicar esa variabilidad. El contenido de elementos mayores justifica las variaciones de luminiscencia intrínseca, asociada a la estructura cristalográfica de los minerales.

4. CONSIDERACIONES PETROGRÁFICAS DE LOS CARBONATOS DE LA FORMACIÓN RIBOTA MEDIANTE CL.

Desde un punto de vista descriptivo, el estudio petrográfico mediante CL permite destacar algunos aspectos de interés, que en un análisis petrográfico convencional hubieran pasado desapercibidos.

No se aprecian variaciones significativas en la respuesta a la CL de los carbonatos de muro a techo de las columnas, ni entre columnas, el rasgo general es el comportamiento homogéneo de los materiales frente a la CL. Esta uniformidad de la luminiscencia en todo el carbonato puede explicarse de dos maneras: puede deberse a que el proceso de dolomitización ha sido muy similar y monofásico, en todas las muestras de la Formación Ribota para diferentes áreas, como correspondería al modelo de dolomitización por enterramiento propuesto en Zamora et al, 1992; o bien, se debe a que el equilibrado diagenético de los contenidos de los elementos Fe y Mn en la red de los carbonatos, durante los procesos de recristalización, no permite discriminar la posible existencia de varios procesos o fases de dolomitización.

Se constata la presencia de texturas primarias relictas, como fragmentos de fósiles y restos de ooides. Dichos relictos aparecen con una luminiscencia mayor que el carbonato que lo rodea o bien no son luminiscentes, esta última opción es porcentualmente la más frecuente.

Hay que resaltar que los restos de ooides aparecen corroídos por la dolomita que los rodea, lo que indicaría que el proceso de dolomitización fagocita parte del resto. Este dato textural es consistente con un proceso de dolomitización tardío y es un criterio más a considerar en la elaboración del modelo de dolomitización (Zamora, 1991 y Zamora *et al.*, 1992).

Otro hecho a tener en cuenta es la influencia del contenido en Fe, como factor modificador del comportamiento del carbonato ante el bombardeo de electrones. Los cristales dolomíticos más luminiscentes son los de textura esparítica y aspecto anubarrado, que presentan evidencia petrográfica de exolución del Fe de la red del carbonato; es decir, con menor contenido de Fe en la estructura del cristal.

También se aprecia una luminiscencia mayor en los cristales de carbonato próximos a superficies estilolíticas, explicable quizás porque asociada a esas superficies de disolución hay una mayor movilidad y perdida del hierro de los carbonatos.

Las principales variaciones de la luminiscencia se observan en los carbonatos de cementación [illegible]tas y venas. Pudiéndose destacar cementos de carbonato con una intensa lun[illegible]cencia naranja, que corresponderían a cementos de calcita, y en muchos casos con bandas o zonaciones de crecimiento en los cristales que reflejan los cambios en la concentración de Mn presente en la red de los cristales de calcita. Esta variación en la concentración se puede explicar de dos formas, por modificaciones en la composición de los elementos traza en el fluido diagenético como consecuencias de modificaciones en el pH y Eh del mismo, o bien por variación en la tasa de crecimiento del cristal. Por el contrario, hay carbonatos no luminiscentes, que normalmente están asociados a fracturas poco definidas a nivel microscópico, tanto en nicoles cruzados como paralelos.

Excepcionalmente se ha detectado la presencia de una luminiscencia azulada en algunos carbonatos, similar a la CL intrínseca, que es la luminiscencia de base de todos los carbonatos y depende exclusivamente de su red cristalográfica (Amieux, 1982). En muchos de los casos estudiados este fenómeno parece deberse a efectos mecánicos de preparación de la lámina, que genera huecos que son rellenados de carbonato procedente del pulido, lo que unido a la luminiscencia azulada de los epóxidos utilizados para la realización de las láminas justificarían esta aparente luminiscencia con tonalidades anómalas en los carbonatos no luminiscentes. No hay que desdeñar tampoco que las rocas analizadas presentan a menudo cuarzos autigénicos con luminiscencia azul, y en algunas muestras existe una luminiscencia azulada de fondo que se debe a la alta silicificación sufrida por las rocas. Sin embargo, en otros casos se observa claramente que es el propio carbonato esparítico el que responde con una ligera luminiscencia en azul. La explicación de este hecho llevaría un estudio paralelo con microsonda electrónica o microscopía electrónica con EDAX (no disponible por ahora) que permitiera observar cuáles son las composiciones de otros elementos traza potenciadores o inhibidores de la luminiscencia de los cristales de carbonato; aunque hay que indicar que existen antecedentes de este tipo de luminiscencia descritos por Sippel y Glover (1965).

La CL nos ha permitido también detectar algún proceso de dedolomitización en vacuolas o áreas de disolución de la roca. Se manifiesta por la presencia de áreas irregulares o bandas de luminiscencia intensa amarillenta, similar a la de las rocas de composición calcítica, seudomorfizando cristales de dolomita con la característica luminiscencia tenue, rojiza.

AGRADECIMIENTOS

Durante la realización de este trabajo A.J. Zamora disfrutó de una Beca de Investigación del Convenio de Colaboración Ibercaja-Universidad de Zaragoza, convocatoria de 1990-91, y de una ayuda del Centro de Estudios Borjanos, XIII Convocatoria.

BIBLIOGRAFÍA

Amieux, P. (1982): La cathodoluminiscence: Méthode d'étude sédimentologique des carbonates. *Bull. Centres Rech. Explor.-Prod. Elf-Aquitaine*, **6**, 437-483.

Fairchild, I.J. (1983): Chemical controls of cathodoluminiscence of natural dolomites and calcites: new data and review. *Sedimentology*, **30**, 579-583.

Have, T. and Heijnen, W. (1985): Cathodoluminiscence activation and zonation in carbonate rocks: an experimental approach. *Geol. Mijnb.*, **64**, 297-310.

Marshall, D. (1988): *Cathodoluminiscence of Geological Materials*. Ed. Unwin Hyman, 146 pp.

Oglesby, T.W. (1976): *A model for the distribution of manganese, iron and magnesium in authigenic calcite and dolomite cements in the upper Smackover Formation in eastern Mississippi*. Masters thesis Univ. Missouri-Columbia, 112 pp.

Pierson, B.J. (1981): The control of cathodoluminiscence in dolomite by iron and manganese. *Sedimentology*, **28**, 601-610.

Sippe, R.F. and Glover, E.D. (1965): Structures in carbonate rocks made visible by luminiscence petrography. *Science*, **150**, 283-287.

Zamora, A. (1991): *Caracterización petrológica de los materiales carbonatados de la Formación Ribota en el ámbito de la Cadena Ibérica Oriental*. Tesis de Licenciatura, Universidad de Zaragoza, 190 pp. inédita.

Zamora A., Mandado, J., Tena, J.M., Auqué, L.F. y Gimeno, M.J. (1992): La Formación Ribota en el Sector Noroeste de la Cadena Ibérica Oriental. Modelo petrogenético. *III Congreso Geológico de España*, Actas tomo 2: 117-121.

Rev. Academia de Ciencias. Zaragoza. 48 (1993)

TRAVERTINOS: REVISIÓN DE LA TERMINOLOGÍA Y CRITERIOS DE CLASIFICACIÓN.

A.J. Zamora, J. Mandado, J.M. Tena y L.F. Auqué

Departamento de Geología. Facultad de Ciencias. Universidad de Zaragoza. 50009 Zaragoza.

ABSTRACT

Terms travertine and calcareous tuff, in a broad sense, refer to continental carbonate accumulations of difficult differentiation. Common criteria to discriminate between those terms (lithification degree, carbonate genesis and water temperature) don't allow to establish a clear definition of both carbonate deposits.

Taken into account genetical, etymological and practical considerations, and to prevent confusion in the terminology, we propose that term "travertine" should be used to designate the incrustations of continental carbonate with abundant vegetal rests and formed by superficial waters.

Travertine carbonate can be strictly generated by physicochemical processes (CO_2 outgassing, evaporation or waters mixing) as wells as by biochemical enhaced precipitation mechanisms.

Two travertine classification schemes can be established: a descriptive classification, mainly based on the type and proportions of organic remains, and a genetic classification based on textural relationship among different components of the carbonate deposits.

1. INTRODUCCIÓN

Los depósitos carbonatados naturales de carácter continental, muestran gran variedad en lo que respecta al medio ambiente donde se originan, diferenciándose cuatro tipos fundamentales (Figura 1): travertino o toba calcárea, carbonatos lacustres, espeleotemas y caliches o costras calcáreas.

De esos cuatro tipos, el término travertino o toba calcárea designa a los carbonatos continentales generados en aguas superficiales corrientes (ríos y surgencias) y en menor medida en lagos, que además conservan abundantes señales de vegetales (micro y macrofitas).

Los antecedentes bibliográficos referentes a los travertinos se remontan a estas dos últimas décadas. Inicialmente el estudio fue enfocado a los aspectos paleontológicos (paleobotánicos) de los mismos, merced al elevado contenido y buena conservación de los restos fósiles, destacando los trabajos de Irion y Müller (1968) y Lang y Lucas (1970). Posteriormente, los trabajos se diversifican en lo que concierne a la temática de estudio; así, basándose en la relación estrecha existente entre los depósitos travertínicos y la climatología del Cuaternario (sobre todo en las regiones mediterráneas y templadas de Europa), que se manifiesta en la necesidad de condiciones húmedas y

de calma tectónica (períodos de biostasia) para la formación de dichos depósitos, los investigadores que trabajan en esta línea realizan trabajos de reconstrucción paleoclimática, destacando la síntesis que aparece en el monográfico de la revista *Mediterranée*, en 1986. Otra temática desarrollada sobre los travertinos consiste en el estudio de las facies (microscópicas y macroscópicas) y sus relaciones laterales y verticales, para el establecimiento de diferentes modelos sedimentológicos y ambientales capaces de generar esos rasgos; en este sentido son básicos los trabajos de Ordóñez y García (1983), Chafetz y Folk (1984), Ferreri y D'Argenio (1985), Ordóñez *et al.* (1986), Pedley (1990) y Lang *et al.* (1992). Por último, hay investigadores como Jacobson y Usdowski (1975), Dandurand *et al.* (1982) y Herman y Lorah (1987 y 1988), entre otros, que consideran a los sistemas travertínicos como laboratorios naturales para el estudio geoquímico del sistema carbonatado; centrando el estudio en aspectos tales como la fisicoquímica y cinética de la precipitación del carbonato, la distribución de elementos menores y traza en la estructura cristalina del precipitado, y el fraccionamiento isotópico del oxígeno y carbono.

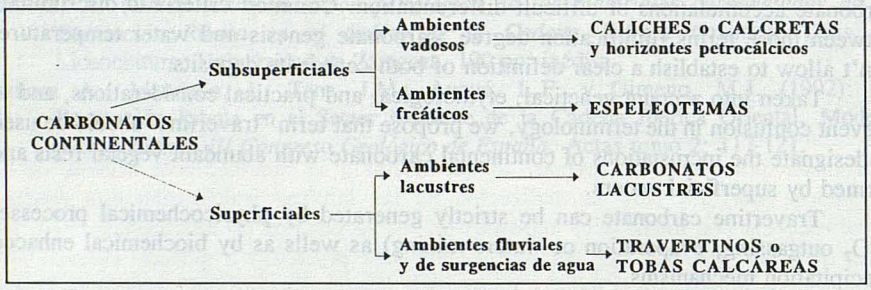


Figura 1.- Cuadro de división de los carbonatos continentales en función del medio genético.

2. TERMINOLOGÍA: ¿ TOBA O TRAVERTINO ?

La utilización de estos dos términos, que hacen referencia según la definición planteada inicialmente al mismo material, está sujeta a una alta confusión debido a que la diferencia básica radica en criterios referidos al grado de litificación (Choppy, 1981; Guendon y Vaudour, 1981; Freytet y Plet, 1991), como son la dureza o friabilidad, porosidad y el carácter pulverulento. En este sentido es de destacar que los carbonatos más modernos reciben el nombre de tobas y, por el contrario, los más antiguos son denominados travertinos.

Algunos autores señalan otras diferencias entre tobas y travertinos, así Füchtbauer (1974) establece que el origen del carbonato en los travertinos sería de tipo inorgánico (abiótico) y, en las tobas, de tipo bioinducido por la actividad biológica de bacterias cianofíceas, musgos y plantas higrófilas. Esta diferencia tiene relación con la asimilación por parte de un gran número de investigadores de habla inglesa (británicos y americanos) del término de travertino al de calc-sinter, sinter o calcareous sinter. Otra diferencia, señalada por Pedley (1990), corresponde a la temperatura del agua a partir de la cual se forman los carbonatos; así, el término travertino se aplica a los carbonatos generados a partir de aguas hidrotermales y el de toba sólo a los de aguas frías. Esta diferencia creemos que proviene de la etimología de la palabra travertino, que deriva del latín "lapis tiburtinus" (piedra de Tibur, actualmente Tívoli), acuñada

para los depósitos carbonatados de esta región que están generados por aguas hidrotermales.

De esta forma, englobando en el concepto inicial estas diferencias deberíamos definir *toba* como un carbonato de origen continental depositado en un ambiente superficial (de ríos, lagos o surgencias), generado por procesos principalmente biológicos (biosíntesis) a partir de aguas frías, caracterizado por su alta porosidad (aspecto esponjoso), friabilidad (blando y frecuentemente pulverulento), por la abundancia de señales de restos vegetales y un contenido variable de detríticos. Por el contrario, *travertino* designa a los carbonatos continentales originados en un ambiente superficial (de ríos, lagos y surgencias), a partir de aguas hidrotermales y por procesos estrictamente fisicoquímicos (abióticos), caracterizado por su buena compactación y dureza (buena litificación), y en el que también se reconoce improntas vegetales.

A pesar de las diferencias apuntadas entre toba y travertino, consideramos, al igual que otros investigadores como Buccino et al. (1978), Julia (1983) y López y Martínez (1989) que debería utilizarse únicamente el término travertino. Elección que basamos en los siguientes argumentos:

- El grado de litificación es un parámetro postdepósito que tiene un efecto variable en intensidad no sólo para depósitos travertínicos de diferentes edades, sino también para un mismo depósito en función de sus características texturales. Además, no tiene relación alguna con la posible génesis diferencial de toba y travertino, sino con el aspecto del carbonato. Estas razones nos hacen descartar la litificación como un rasgo distintivo entre ambos.

- En lo que respecta al diferente origen del carbonato para uno y otro término, consideramos que no es un carácter debidamente contrastado, por la problemática todavía existente sobre el papel de los seres vivos en la precipitación del carbonato.

- La temperatura del agua es una distinción que únicamente se puede establecer a priori en sistemas actuales de génesis de travertinos, por el contrario en medios no funcionales (son mayoritarios) no se puede determinar inicialmente y por tanto no es un parámetro operativo de distinción.

- Desde el punto de vista etimológico, el término toba proviene del latín "tufa" y este a su vez del griego "tophus", y simplemente hace alusión a la friabilidad de la roca por lo que no puede aplicarse estrictamente a este tipo de carbonatos; por el contrario, el término travertino se acuñó originariamente en Italia para este tipo de materiales.

- La existencia de un doble significado en el término toba, que hace referencia no sólo a lo carbonatos continentales sino también a materiales volcanoclásticos.

En conclusión, el concepto que vamos a utilizar para el término travertino, basándonos en Buccino *et al.* (1978), Chafetz y Folk (1982) y Julia (1983) es el siguiente:

"El término travertino designa un depósito continental con un variable grado de litificación y básicamente constituido por carbonato cálcico; originado en un medio continental superficial de aguas dulces (frías o hidrotermales) asociadas a surgencias de agua, cauces fluviales y zonas lacustres; formado por incrustación de tipo fisicoquímico o bioquímico (bioinducido) sobre soportes vegetales vivos o muertos (de microfitas y macrofitas) u otro tipo de soporte, e incluyendo además las acumulaciones clásticas originadas por destrucción de travertinos previos; y en el que prevalece la estructura vegetal".

De esta definición hay que destacar dos aspectos nuevos, que se han tenido en cuenta:

- El carácter incrustante del carbonato, que precipita sobre una superficie a la que recubre, independientemente de la naturaleza de la misma, quedando incluidos dentro del término travertino no sólo los depósitos formados sobre un substrato vegetal, sino también aquellos que aparecen sobre un objeto o substrato inorgánico.
- La consideración expresa, dentro del término travertino, de aquellas acumulaciones de fragmentos de travertinos previos (fitoclastos) de tamaño variable, originados por destrucción de los mismos.

3. GÉNESIS DEL CARBONATO EN LOS TRAVERTINOS

Actualmente existe una problemática abierta sobre el origen del carbonato en los travertinos, que se plantea en los siguientes términos: ¿Cuál es la influencia o papel que desarrollan los seres vivos en la precipitación del carbonato que constituye los travertinos?.

Así, existen investigadores, como Bouyx y Pias (1971), Lorah y Herman (1988), Viles y Goudie (1990) y Pentecost (1990), que señalan que el carbonato precipita debido fundamentalmente a la acción de procesos de tipo fisicoquímico. Otros, como Golubic (1969), Casanova (1981) y Ordóñez *et al.* (1986), por el contrario, reconocen la influencia de los procesos biológicos en la precipitación del carbonato, pero le asignan un carácter variable en función de la energía del medio, de tal forma que en ambientes de alta energía (cabeceras de los ríos, zonas de cascadas y rápidos), son los procesos inorgánicos los más influyentes, y en ambientes de energía baja (tramos bajos del río y zonas aguas arriba de represas travertínicas) son los procesos biológicos la causa principal de la precipitación.

Para los partidarios de establecer la acción de los seres vivos como causa generadora de los depósitos travertínicos, se plantea una disyuntiva entre considerar la actividad algal como la principal causante de la precipitación (Guendon y Vaudour, 1981); o bien, la actividad bacteriana (Adolphe 1981 y Adolphe *et al.* 1989).

Al margen de esta problemática, podemos agrupar en dos los diferentes procesos generadores que han sido establecidos para explicar la precipitación del carbonato en travertinos: Los procesos biológicos y los fisicoquímicos (también denominados inorgánicos o abióticos.

En los **procesos biológicos,** los organismos vivos, principalmente bacterias y algas, son parte activa en la precipitación del carbonato por modificación de los parámetros que regulan el sistema carbonatado, denominándose a este proceso como precipitación bioinducida o biosíntesis; o bien, realizan un papel pasivo de atrapamiento mecánico de partículas.

Para el caso de la precipitación bioquímica del carbonato, se han señalado varias actividades biológicas que pueden causar una modificación en los equilibrios carbonatados: la fotosíntesis, principalmente de algas, y la actividad bacteriana sobre la materia orgánica. En cuanto al primero, su efecto sobre el sistema carbonatado consiste en una degasificación o pérdida de CO_2 debido al consumo que del mismo realizan las plantas con clorofila para sintetizar la materia orgánica. Esta pérdida de CO_2 repercute en un aumento de la basicidad del medio y en que puedan llegar a alcanzarse condiciones de sobresaturación de las aguas en carbonato cálcico. En lo que respecta a la actividad bacteriana sobre la materia orgánica, su efecto sobre los

equilibrios carbonatados depende del grado de oxigenación del medio. Así, en un medio oxigenado (medios aerobios), la oxidación de los compuestos de carbono y nitrógeno constitutivos de la materia orgánica produce un aumento del CO_2 disuelto y un descenso del pH, y consecuentemente una tendencia a desplazar los equilibrios de precipitación-disolución del carbonato cálcico hacia la disolución. Por el contrario, en medios anaerobios, el efecto de reducción de la materia orgánica sobre los equilibrios carbonatados es más complejo, de tal manera que las reacciones que afectan a los compuestos de carbono producen un aumento de la acidez del medio, mientras que las relacionadas con el proceso de desnitrificación de los compuestos orgánicos nitrogenados producen un incremento de la basicidad del mismo.

En el grupo de los **procesos fisicoquímicos** se incluyen aquellos fenómenos que, sin intervención alguna de los seres vivos, producen la modificación de los equilibrios carbonatados.

Entre los procesos posibles se citan la desgasificación del CO_2 y, en menor medida, la concentración evaporativa y la mezcla de aguas (Flügel, 1982). En lo que respecta al proceso de desgasificación, hay varios mecanismos que pueden generar la pérdida de CO_2:

- Por turbulencia en el agua, mecanismo muy importante en ambientes de cascadas travertínicas (Ordóñez *et al.*, 1979; Lorah y Herman, 1988). El fenómeno consiste básicamente en que el efecto de la turbulencia produce una mayor aireación del agua y facilita de esta manera la pérdida de CO_2 de las aguas, que en flujos laminares o aguas estancadas permanecería disuelto.
- Por diferencias en la presión y temperatura entre el agua y el aire. Este mecanismo es importante en los puntos de surgencia de aguas subterráneas, donde se produce un desequilibrio entre las condiciones de las aguas surgentes y el entorno, manifestado por una variación en la presión y temperatura; así, los descensos de presión e incrementos de temperatura favorecen el proceso de desgasificación y la subsiguiente precipitación del carbonato.

La influencia del proceso de evaporación en la formación de travertinos aumenta para el caso de aguas poco móviles en medios de aridez media a elevada, y consiste básicamente en la sobresaturación del medio respecto al carbonato cálcico por efecto de la perdida del disolvente (agua) debida a la evaporación.

El proceso de mezcla de aguas es otro de los mecanismos citados para la génesis del carbonato de los travertinos, si bien su importancia es muy reducida. Sólo en casos muy específicos, cuando dos aguas de composición diferente (por ejemplo un cauce de agua que atraviesa una zona de yesos y otro que atraviesa una zona calcárea), pero que tienen un ión común (en nuestro ejemplo el calcio), se mezclan, produciéndose la sobresaturación en este catión y la precipitación de carbonato cálcico.

4. CLASIFICACIÓN DE TRAVERTINOS

En la sistemática de los travertinos hay una gran cantidad de clasificaciones y términos asociados. Todas ellas las podemos agrupar en dos tipos básicos: aquellas que se aplican a un depósito travertínico o sistema travertínico, entendiendo como tal el conjunto de facies asociadas a un determinado ambiente o medio sedimentario (escala megascópica); y las que se aplican a un travertino; es decir, a la roca vista a escala de visu (escala macroscópica).

En lo que concierne a las clasificaciones establecidas para los depósitos travertínicos, todas ellas utilizan como criterio de subdivisión el medio ambiente o entorno geomorfológico donde se producen estos depósitos. Según Pedley (1990), hay tres tipos de modelos ambientales (equivalente al concepto de sistema travertínico):

- Travertinos de vertiente; se asocian a surgencias de agua en laderas de pendiente variable y más o menos alejadas del cauce fluvial.
- Travertinos fluviales, que constituyen la mayoría de los depósitos travertínicos conocidos y como su nombre indica se generan en la zona de influencia directa de los cauces fluviales.
- Travertinos lacustres, que corresponden a los depósitos originados en las zonas someras de lagos que aparecen colonizados por macro y microfitas.

Las clasificaciones aplicadas a los travertinos (escala macroscópica) se agrupan en dos tipos en función del criterio de subdivisión de las mismas:

a) Composicionales o descriptivas. En ellas se considera únicamente la naturaleza y abundancia de los restos fósiles (Irion y Müller, 1968; Lang y Lucas, 1970), o bien la morfología del encostramiento (Casanova, 1981).

b) Texturales o interpretativas. Cuando se utiliza como criterio de subdivisión el carácter transportado o in situ de la incrustación, y la relación entre los distintos componentes que constituyen el travertino (Ferreri y D'Argenio, 1985; Pedley, 1990).

Las clasificaciones basadas en el contenido de restos fósiles son de tipo cualitativo y originariamente son las que primero aparecen debido al interés paleontológico de los travertinos. La utilización de las mismas ha supuesto la aparición de gran cantidad de términos, caracterizados únicamente por la presencia de un grupo florístico o faunístico abundante o especial. Entre ellos podemos citar los de: travertinos de algas, travertinos de musgos o briofitas, travertinos de helechos, travertinos de gasterópodos y travertinos de chironomides; en los que el componente fósil destacado es, respectivamente, las algas, musgos, helechos, gasterópodos y tubos de larvas de insectos.

Las clasificaciones texturales suponen un paso hacia adelante en lo que respecta a la interpretación sedimentológica y paleoambiental de los travertinos. En ellas aparece un primer criterio de subdivisión que trata sobre el origen de la incrustación, diferenciando entre los travertinos autóctonos, en los cuales el carbonato precipitado conserva la estructura vegetal en la posición originaria, es decir, la incrustación permanece in situ; y los travertinos alóctonos o clásticos que están constituidos por restos fósiles individualizados (hojas y tallos) acumulados y posteriormente incrustados por el carbonato o bien por fragmentos travertínicos resultantes de la destrucción de travertinos previos, presentando todos ellos una textura clástica.

Las subdivisiones de estos dos grupos es variable; así, Pedley (1990) en los carbonatos autóctonos, utilizando un criterio similar al de Embry y Klovan (1971) para los carbonatos marinos bioconstruidos, diferencia entre framestone fitohermal, que corresponde a los travertinos constituidos por plantas higrófilas en posición de vida, que constituyen el esqueleto o armazón de la roca, y boundstone fitohermal, término equivalente al concepto de estromatolito continental. Ferreri y D'Argenio (1985), además de esos dos tipos, a los que denominan respectivamente travertinos fitohermales y estromatolíticos, incluyen los travertinos microhermales, que corresponden a la incrustación in situ sobre plantas higrófilas de tamaño pequeño (musgos y helechos).

La subdivisión establecida por Pedley (1990) para los carbonatos clásticos es cualitativa, y se basa en la naturaleza del grano o clasto. Este autor diferencia entre

travertinos fitoclásticos (los que están constituidos por fragmentos de plantas individualizados -hojas o tallos- que durante o posteriormente a su transporte han sido incrustados por carbonato), travertinos cianolíticos (constituidos principalmente por oncolitos), travertinos intraclásticos (los formados por fragmentos de travertinos previos que han sido destruidos mecánicamente) y travertinos microdetríticos. Ferreri y D'Argenio (1985), por el contrario, utilizan criterios más cuantitativos y muy similares a los de Dunham (1962) para los carbonatos detríticos marinos.

AGRADECIMIENTOS

Durante la realización de este trabajo A. J. Zamora ha disfrutado de una Beca de Investigación del Consejo Asesor de Investigación de la Diputación General de Aragón (CONAI, BCB-12/91).

BIBLIOGRAFIA

Adolphe, J.P. (1981): *Assoc. Fr. Karstologie*, 3, 15-30.

Adolphe, J.P.; Hourimèche, A.; Loubière, A.; Loubière, J.F.; Paradas, J. et Soleilhavoup, F. (1989): *Bull. Soc. Geol. France*, V (1), 55-62.

Bouyx, E. et Pias, J. (1971): *C. R. Acad. Sci. Paris*, 273, 2468-2471.

Buccino, G.; D'Argenio, B.; Ferreri, V.; Brancaccio, L.; Ferreri, M.; Panichi, C. & Stanzione, D. (1978): *Boll. Soc. Geol. It.*, 97, 617-646.

Casanova, J. (1981): *Assoc. Fr. Karstologie*, mem. 3, 45-54.

Chafetz, H.S. & Folk, R.L. (1984): *Jour. Sed. Petrol.*, 54 (1), 289-316.

Choppy, J. (1981): *Assoc. Fr. Karstologie*, mem. 3, 55-60.

Dandurand, J.L.; Gout, R.; Hoefs, J.; Menschel, G.; Schott, J. & Usdowski, E. (1982): *Chem. Geol.*, 36, 299-315.

Dunham, R.J. (1962): *Amer. Assoc. Petrol. Geol.*, Symp., 108-121.

Embry, A.F. & Klovan, J.E. (1971): *Can. Petrol. Geol. Bull.*, 19, 730-781.

Ferreri, V. e D'Argenio, B. (1985): *Rend. Acc. Sc. Fis. e Matem.*, IV (LII/2), 1-47.

Flügel, E. (1982): *Microfacies analysis of limestones*. Springer-Verlag. Berlin.

Freytet, P. & Plet, A. (1991): *Geobios*, 24 (2), 123-139.

Füchtbauer, H. (1974): *Sediments and sedimentary rocks*. Schweizerbart. Stuttgart.

Golubic, S. (1969): *Verh. Inter. Verein. Limnol.*, 17, 956-961.

Guendon, J.L. et Vaudour, J. (1981): *Assoc. Fr. Karstologie*, mem. 3, 89-100.

Herman, J.S. & Lorah, M.M. (1987): *Chem. Geol.*, 62, 251-262.

Herman, J.S. & Lorah, M.M. (1988): *Geochim. Cosmochim. Acta*, 52, 2347-2355.

Irion, G. & Müller, G. (1968): *Recent developments in carbonate sedimentology in central Europe*. Springer-Verlag. Berlin.

Jacobson, R.L. & Usdowski, E. (1975): *Contrib. Mineral. Petrol.*, 51, 65-74.

Julia, R. (1983): *Amer. Assoc. Petrol. Geol.*, mem. 3, 64-72.

Lang, J. et Lucas, G. (1970): *Bull. Soc. Geol. France*, 7 (XII), 634-642.

Lang, J.; Pascal, A. & Salomon, J. (1992): *Z. Geomorph. N.F.*, 36 (3), 273-291.

López, F. y Martínez, J. (1989): *Bol. Geol. Min.*, 100, 248-258.

Lorah, M.M. & Herman, J.S. (1988): *Water Resources Res.*, 24 (9), 1541-1552.

Ordóñez, S. & García Del Cura, M.A. (1983): *Spec. Publ. Int. Ass. Sediment.*, 6, 485-497.

Ordóñez, S.; González, J.A. y García Del Cura, M.A. (1979): *IV Reun. Grup. Trabajo Cuaternario*, 171-178.

Ordóñez, S.; González, J.A. y García Del Cura, M.A. (1986): *Rev. Mat. Proc. Geol.*, IV, 229-255.

Pedley, H.M. (1990): *Sedim. Geol.*, 68, 143-154.

Viles, H.A. & Goudie, A.S. (1990): *Earth Surface Processes and Landforms*, 15, 425-443.