An Invitation to Smooth Infinitesimal Analysis

John L. Bell

In the usual development of the calculus, for any differentiable function f on the real line **R**, $y = f(x)$, it follows from Taylor's theorem that the increment $\delta y = f(x + \delta x) - f(x)$ in y attendant upon an increment δx in x is determined by an equation of the form

$$\delta y = f'(x)\delta x + A(\delta x)^2, \qquad (1)$$

where $f'(x)$ is the derivative of $f(x)$ and A is a quantity whose value depends on both x and δx. Now if it were possible to take δx so *small* (but not demonstrably identical with 0) that $(\delta x)^2 = 0$ then (1) would assume the simple form

$$f(x + \delta x) - f(x) = \delta y = f'(x)\,\delta x. \qquad (2)$$

We shall call a quantity having the property that its square is zero a *nilsquare infinitesimal* or simply an *infinitesimal.* In *smooth infinitesimal analysis* (*SIA*)[1] "enough" infinitesimals are present to ensure that equation (2) holds *nontrivially* for *arbitrary* functions $f: \mathbf{R} \to \mathbf{R}$. (Of course (2) holds trivially in standard mathematical analysis because there 0 is the sole infinitesimal in this sense.) The meaning of the term "nontrivial" here may be explicated in following way. If we replace δx by the letter ε standing for an arbitrary infinitesimal, (2) assumes the form

$$f(x + \varepsilon) - f(x) = \varepsilon f'(x). \qquad (3)$$

Ideally, we want the validity of this equation to be independent of ε , that is, given x, for it to hold for *all* infinitesimal ε. In that case the derivative $f'(x)$ may be *defined* as the unique quantity D such that the equation

$$f(x + \varepsilon) - f(x) = \varepsilon D$$

holds for all infinitesimal ε.

Setting $x = 0$ in this equation, we get in particular

$$f(\varepsilon) = f(0) + \varepsilon D, \qquad (4)$$

for all ε. *It is equation* (4) *that is taken as axiomatic in smooth infinitesimal*

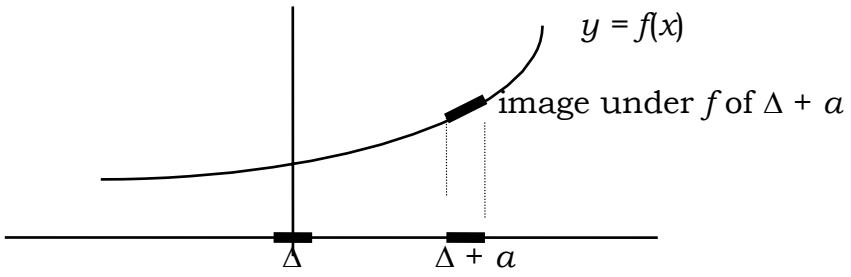
[1] For a detailed development of smooth infinitesimal analysis, see Bell [1998].

analysis. Let us write Δ for the set of infinitesimals, that is,

$$\Delta = \{x\colon x \in \mathbf{R} \wedge x^2 = 0\}.$$

Then it is postulated that, for any $f\colon \Delta \to \mathbf{R}$, there is a *unique* $D \in \mathbf{R}$ such that equation (4) holds for all ε. This says that the graph of f is a straight line passing through $(0, f(0))$ with slope D. Thus any function on Δ is what mathematicians term *affine*, and so this postulate is naturally termed the *principle of infinitesimal affineness,* of *microstraightness.* It means that Δ *cannot be bent or broken*: it is subject only to *translations and rotations*—and yet is not (as it would have to be in ordinary analysis) identical with a point. Δ may be thought of as an entity possessing position and attitude, but lacking true extension.

If we think of a function $y = f(x)$ as defining a curve, then, for any a, the image under f of the "infinitesimal interval" $\Delta + a$ obtained by translating Δ to a is straight and coincides with the tangent to the curve at $x = a$ (see figure immediately below). In this sense each curve is "infinitesimally straight".



From the principle of infinitesimal affineness we deduce the important *principle of infinitesimal cancellation*, viz.

$$\text{IF } \varepsilon a = \varepsilon b \textit{ FOR ALL } \varepsilon, \text{ THEN } a = b.$$

For the premise asserts that the graph of the function $g\colon \Delta \to \mathbf{R}$ defined by $g(\varepsilon) = a\varepsilon$ has both slope a and slope b: the uniqueness condition in the principle of infinitesimal affineness then gives $a = b$. The principle of infinitesimal cancellation supplies the exact sense in which there are "enough" infinitesimals in smooth infinitesimal analysis.

From the principle of infinitesimal affineness it also follows that *all functions on* $\mathbf{R}$ *are continuous*, that is, *send neighbouring points to neighbouring points.* Here two points x, y on $\mathbf{R}$ are said to be neighbours if $x - y$ is in Δ, that is, if x and y differ by an infinitesimal. To see this, given $f\colon \mathbf{R} \to \mathbf{R}$ and

neighbouring points x, y, note that $y = x + \varepsilon$ with ε in Δ , so that

$$f(y) - f(x) = f(x + \varepsilon) - f(x) = \varepsilon f'(x).$$

But clearly any multiple of an infinitesimal is also an infinitesimal, so $\varepsilon f'(x)$ is infinitesimal, and the result follows.

In fact, since equation (3) holds for any f, it also holds for its derivative f'; it follows that functions in smooth infinitesimal analysis are differentiable arbitrarily many times, thereby justifying the use of the term "smooth".

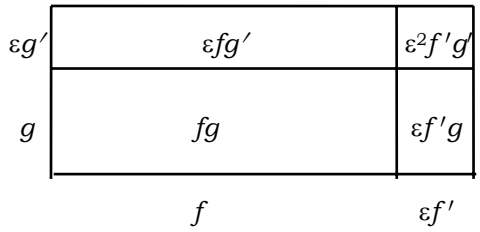
Let us derive a basic law of the differential calculus, the *product rule:*

$$(fg)' = f'g + fg'.$$

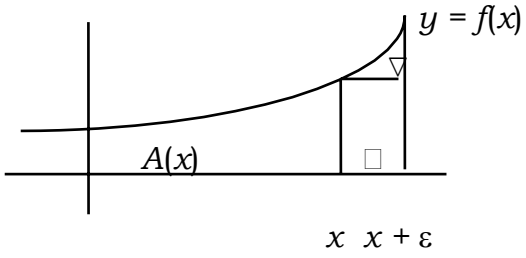
To do this we compute

$$\begin{aligned}(fg)(x + \varepsilon) &= (fg)(x) + (fg)'(x) = f(x)g(x) + (fg)'(x),\\ (fg)(x + \varepsilon) &= f(x + \varepsilon)g(x + \varepsilon) = [f(x) + f'(x)].[g(x) + g'(x)]\\ &= f(x)g(x) + \varepsilon(f'g + fg') + \varepsilon^2 f'g'\\ &= f(x)g(x) + \varepsilon(f'g + fg'),\end{aligned}$$

since $\varepsilon^2 = 0$. Therefore $\varepsilon(fg)' = \varepsilon(f'g + fg')$, and the result follows by infinitesimal cancellation. This calculation is depicted in the diagram below.



Next, we derive the *Fundamental Theorem of the Calculus*.



Let J be a closed interval $\{x\colon a \leq x \leq b\}$ in **R** and $f\colon J \to$ **R**; let $A(x)$ be the

area under the curve $y = f(x)$ as indicated above. Then, using equation (3),

$$\varepsilon A'(x) = A(x + \varepsilon) - A(x) = \square + \triangledown = \varepsilon f(x) + \triangledown.$$

Now by infinitesimal affineness $\triangledown$ is a triangle of area $\frac{1}{2}\varepsilon.\varepsilon f'(x) = 0$. Hence $\varepsilon A'(x) = \varepsilon f(x)$, so that, by infinitesimal cancellation,

$$A'(x) = f(x).$$

A *stationary point* a in $\mathbf{R}$ of a function $f: \mathbf{R} \to \mathbf{R}$ is defined to be one in whose vicinity "infinitesimal variations" fail to change the value of f, that is, such that $f(a + \varepsilon) = f(a)$ for all ε. This means that $f(a) + \varepsilon f'(a) = f(a)$, so that $\varepsilon f'(a) = 0$ for all ε, whence it follows from infinitesimal cancellation that $f'(a) = 0$. This is *Fermat's rule.*

An important postulate concerning stationary points that we adopt in smooth infinitesimal analysis is the

Constancy Principle. If every point in an interval J is a stationary point of f: $J \to \mathbf{R}$ (that is, if f' is identically 0), then f is constant.

Put succinctly, "universal local constancy implies global constancy". It follows from this that two functions with identical derivatives differ by at most a constant.

In ordinary analysis the continuum $\mathbf{R}$ is connected in the sense that it cannot be split into two non empty subsets neither of which contains a limit point of the other. In smooth infinitesimal analysis it has the vastly stronger property of *indecomposability:* it cannot be split *in any way whatsoever* into two disjoint nonempty subsets. For suppose $\mathbf{R} = U \cup V$ with $U \cap V = \varnothing$. Define $f: \mathbf{R} \to \{0, 1\}$ by $f(x) = 1$ if $x \in U$, $f(x) = 0$ if $x \in V$. We claim that f is constant. For we have

$$(f(x) = 0 \text{ or } f(x) = 1) \quad \& \quad (f(x + \varepsilon) = 0 \text{ or } f(x + \varepsilon) = 1).$$

This gives 4 possibilities:

(i) $f(x) = 0 \quad \& \quad f(x + \varepsilon) = 0$
(ii) $f(x) = 0 \quad \& \quad f(x + \varepsilon) = 1$
(iii) $f(x) = 1 \quad \& \quad f(x + \varepsilon) = 0$
(iv) $f(x) = 1 \quad \& \quad f(x + \varepsilon) = 1$

Possibilities (ii) and (iii) may be ruled out because f is continuous. This leaves (i) and (iv), in either of which $f(x) = f(x + \varepsilon)$. So f is locally, and hence globally, constant, that is, constantly 1 or 0. In the first case $V = \varnothing$, and in the second $U = \varnothing$.

We observe that the postulates of smooth infinitesimal analysis are *incompatible with the law of excluded middle of classical logic*. This incompatibility can be demonstrated in two ways, one informal and the other rigorous. First the informal argument. Consider the function f defined for real numbers x by $f(x) = 1$ if $x = 0$ and $f(x) = 0$ whenever $x \neq 0$. If the law of excluded middle held, each real number would then be either equal or unequal to 0, so that the function f would be defined on the whole of **R.** But, considered as a function with domain **R,** f is clearly discontinuous. Since, as we know, in smooth infinitesimal analysis every function on **R** is continuous, f cannot have domain **R** there[2]. So the law of excluded middle fails in smooth infinitesimal analysis. To put it succinctly, *universal continuity implies the failure of the law of excluded middle.*

Here now is the rigorous argument. We show that the failure of the law of excluded middle can be derived from the principle of infinitesimal cancellation. To begin with, if $x \neq 0$, then $x^2 \neq 0$, so that, if $x^2 = 0$, then necessarily not $x \neq 0$. This means that

$$\textit{for all infinitesimal } \varepsilon, \textit{ not } \varepsilon \neq 0. \qquad (*)$$

Now suppose that the law of excluded middle were to hold. Then we would have, for any ε, either $\varepsilon = 0$ or $\varepsilon \neq 0$. But (*) allows us to eliminate the second alternative, and we infer that, for all ε, $\varepsilon = 0$. This may be written

$$\textit{for all } \varepsilon, \ \ \varepsilon.1 = \varepsilon.0,$$

from which we derive by infinitesimal cancellation the falsehood $1 = 0$. So again the law of excluded middle must fail.

The "internal" logic of smooth infinitesimal analysis is accordingly not full classical logic. It is, instead, *intuitionistic* logic, that is, the logic derived from the constructive interpretation of mathematical assertions. In our brief sketch we did not notice this "change of logic" because, like much of elementary mathematics, the topics we discussed are naturally treated by constructive means such as direct computation.

What are the *algebraic* and *order structures* on **R** in SIA? As far as the former is concerned, there is little difference from the classical situation: in SIA **R** is equipped with the usual addition and multiplication operations under which it is a field. In particular, **R** satisfies the condition that each $x \neq 0$ has a multiplicative inverse. Notice, however, that since in SIA no microquantity

[2] The domain of f is in fact $(\mathbf{R} - \{0\}) \cup \{0\}$, which, because of the failure of the law of excluded middle in SIA, is provably unequal to **R.**

(apart from 0 itself) is provably ≠ 0, microquantities are not required to have multiplicative inverses (a requirement which would lead to inconsistency). From a strictly algebraic standpoint, **R** in SIA differs from its classical counterpart only in being required to satisfy the principle of infinitesimal cancellation.

The situation is different, however, as regards the order structure of **R** in SIA. Because of the failure of the law of excluded middle, the order relation < on **R** in SIA cannot satisfy the trichotomy law

$$x < y \vee y < x \vee x = y,$$

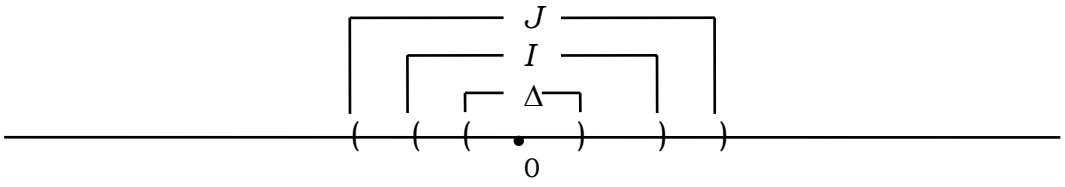
and accordingly < must be a *partial,* rather than a *total* ordering. Since microquantities do not have multiplicative inverses, and **R** is a field, any microquantity ε must satisfy

$$\neg\, \varepsilon < 0 \ \wedge \neg\, \varepsilon < 0.$$

Accordingly, if we define the relation ≤ ("not less than") $x < y$, then, for any microquantity ε we have

$$\varepsilon \leq 0 \wedge \varepsilon \geq 0.$$

Using these ideas we can identify three distinct *infinitesimal neighbourhoods* of 0 on **R** in SIA, each of which is included in its successor. First, the set Δ of



microquantities itself, next, the set $I = \{x \in \mathbf{R}: \neg\ x \neq 0\}$ of elements indistinguishable from 0; finally, the set $J = \{x \in \mathbf{R}: x \leq 0 \wedge x \geq 0\}$ of elements neither less nor greater than 0. These three may be thought of as the infinitesimal neighbourhoods of 0 defined *algebraically, logically, and order-theoretically,* respectively.

In certain models of SIA the system of *natural numbers* possesses some subtle and intriguing features which make it possible to introduce another type of infinitesimal—the so-called *invertible* infinitesimals—resembling those of nonstandard analysis, whose presence engenders yet another infinitesimal

neighbourhood of 0 properly containing all those introduced above.

In SIA the set **N** of natural numbers can be defined to be the smallest subset of **R** which contains 0 and is closed under the operation of adding 1. In some models of SIA, **R** satisfies the *Archimedean principle* that every real number is majorized by a natural number. However, models of SIA have been constructed (see Moerdijk and Reyes [1991]) in which **R** is not Archimedean in this sense. In these models it is more natural to consider, in place of **N**, the set **N*** of *smooth natural numbers* defined by

$$\mathbf{N}^* = \{x \in \mathbf{R}: 0 \leq x \wedge \sin \pi x = 0\}.$$

N* is the set of points of intersection of the smooth curve $y = \sin \pi x$ with the positive x-axis. In these models **R** can be shown to possess the Archimedean property *provided that in the definition* **N** *is replaced by* **N*.** In these models, then, **N** is a proper subset of **N***: the members of **N*** **–** **N** may be considered *nonstandard integers.* Multiplicative inverses of nonstandard integers are infinitesimals, but, being themselves invertible, they are of a different type from the ones we have considered so far. It is quite easy to show that they, as well as the infinitesimals in J (and so also those in Δ and I) are all contained in the set—a further infinitesimal neighbourhood of 0—

$$K = \{x \in \mathbf{R}: \forall n \in \mathbf{N}. -1/n+1 < x < 1/n+1\}$$

of *infinitely small* elements of **R.** The members of the set

$$In = \{x \in K: x \neq 0\}$$

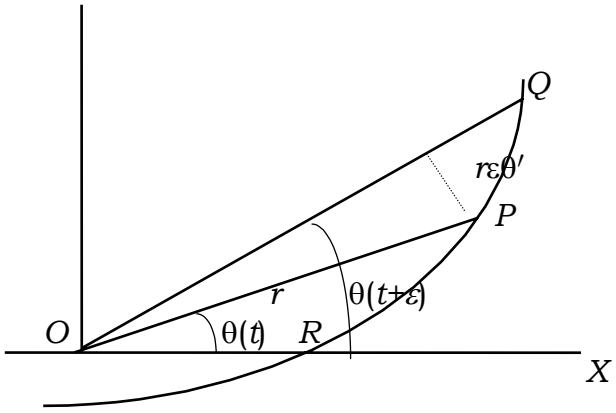
of invertible elements of K are naturally identified as *invertible* infinitesimals. Being obtained as inverses of "infinitely large" reals (i.e. reals r satisfying $\forall n \in \mathbf{N}. n < r \vee \forall n \in \mathbf{N}. r < -n$)the members of In are the counterparts in SIA of the infinitesimals of nonstandard analysis.

*

We conclude with two applications of SIA to physics.

First, we derive the *Kepler-Newton areal law of motion under a central force.* We suppose that a particle executes plane motion under the influence of a force directed towards some fixed point O. If P is a point on the particle's trajectory with coordinates x, y, we write r for the length of the line PO and θ for the angle that it makes with the x-axis OX. Let A be the area of the sector ORP, where R is the point of intersection of the trajectory with OX. We regard x, y, r, θ as functions of a time variable t: thus

$$x = x(t),\ y = y(t),\ r = r(t),\ \theta = \theta(t),\ A = A(t).$$



Now let Q be a point on the trajectory at which the time variable has value $t + \varepsilon$, with ε in Δ. Then by Microstraightness the sector OPQ is a triangle of base $r(t + \varepsilon) = r + \varepsilon r'$ and height

$$r \sin[\theta(t + \varepsilon) - \theta(t)] = r \sin \varepsilon\theta' = r\varepsilon\theta'.$$

(Here we note that $\sin \varepsilon = \varepsilon$ for microquantities ε: recall that $\sin x$ is approximately equal to x for small values of x.) The area of OPQ is accordingly

$$\tfrac{1}{2} \text{ base} \times \text{height} = \tfrac{1}{2} (r + \varepsilon r')r\varepsilon\theta' = \tfrac{1}{2}(r^2\varepsilon\theta' + \varepsilon^2 rr'\theta') = \tfrac{1}{2} r^2\varepsilon\theta'.$$

Therefore

$$\varepsilon A'(t) = A(t + \varepsilon) - A(t) = \text{area } OPQ = \tfrac{1}{2}\varepsilon r^2\theta',$$

so that, cancelling ε,

$$A'(t) = \tfrac{1}{2}r^2\theta'. \qquad (*)$$

Now let $H = H(t)$ be the acceleration towards O induced by the force. Resolving the acceleration along and normal to OX, we have

$$x'' = H\cos\theta \qquad y'' = H\sin\theta.$$

Also $x = r\cos\theta$, $y = r\sin\theta$. Hence

$$yx'' = Hy\cos\theta = Hr\sin\theta\cos\theta \qquad x\,y'' = Hx\sin\theta = Hr\sin\theta\cos\theta,$$

from which we infer that

$$(xy' - yx')' = xy'' - yx'' = 0.$$

Hence

$$xy' - yx' = k, \qquad (**)$$

where k is a constant.

Finally, from $x = r\cos\theta$, $y = r\sin\theta$, it follows in the usual way that

$$xy' - yx' = r^2\theta',$$

and hence, by (**) and (*), that

$$2A'(t) = k.$$

Assuming $A(0) = 0$, we conclude that

$$A(t) = \tfrac{1}{2}kt.$$

Thus the radius vector joining the body to the point of origin sweeps out equal areas in equal times (Kepler's law).

Finall, a remark on the form of *spacetime metrics* in smooth infinitesimal analysis. In a spacetime the metric can be written in the form

(*) $$ds^2 = \Sigma g_{\mu\nu}dx_\mu dx_\nu \qquad \mu,\nu = 1,2,3,4.$$

In the classical setting (*) is in fact an abbreviation for an equation involving derivatives and the "differentials" ds and dx_μ are not really quantities at all. What form does this equation take in SIA? Notice that the "differentials" cannot be taken as nilsquare infinitesimals since all the squared terms would vanish. But the equation does have a very natural form in terms of nilsquare infinitesimals. Here is an informal way of obtaining it.

We think of the dx_μ as being multiples $k_\mu e$ of some small quantity e. Then (*) becomes

$$ds^2 = e^2\Sigma g_{\mu\nu}k_\mu k_\nu,$$

so that

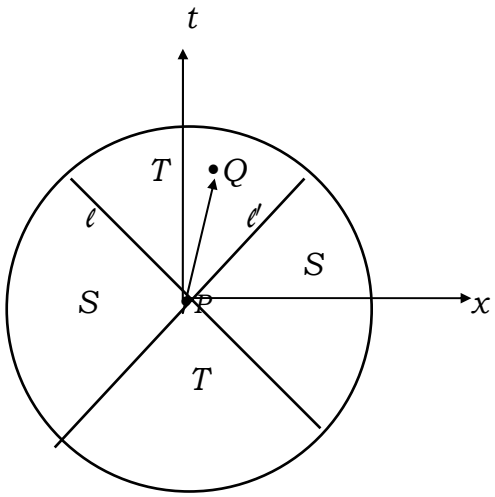
$$ds = e[\Sigma g_{\mu\nu} k_\mu k_\nu]^{1/2}$$

Now replace e by a nilsquare infinitesimal ε. Then we obtain the metric relation in SDG:

$$ds = \varepsilon[\Sigma g_{\mu\nu} k_\mu k_\nu]^{1/2}.$$

This tells us that the "infinitesimal distance" ds between a point P with coordinates (x_1, x_2, x_3, x_4) and an infinitesimally near point Q with coordinates $(x_1 + k_1\varepsilon, x_2 + k_2\varepsilon, x_3 + k_3\varepsilon, x_4 + k_4\varepsilon)$ is $\varepsilon[\Sigma g_{\mu\nu} k_\mu k_\nu]^{1/2}$. Here a curious situation arises. For when the "infinitesimal interval" ds between P and Q is timelike (or lightlike), the quantity $\Sigma g_{\mu\nu} k_\mu k_\nu$ is nonnegative, so that its square root is a real number. In this case ds may be written as εd, where d is a real number. On the other hand, if ds is spacelike, then $\Sigma g_{\mu\nu} k_\mu k_\nu$ is negative, so that its square root is imaginary. In this case, then, ds assumes the form $i\varepsilon d$, where d is a real number (and, of course $i = \sqrt{-1}$). On comparing these we see that, if we take ε as the "infinitesimal unit" for measuring infinitesimal timelike distances, then $i\varepsilon$ serves as the "imaginary infinitesimal unit" for measuring infinitesimal spacelike distances.

For purposes of illustration, let us restrict the spacetime to two dimensions (x, t), and assume that the metric takes the simple form $ds^2 = dt^2 - dx^2$. The infinitesimal light cone at a point P divides the infinitesimal neighbourhood at P into a timelike region T and a spacelike region S,



bounded by the null lines ℓ and ℓ' respectively. If we take P as origin of coordinates, a typical point Q in this neighbourhood will have coordinates $(a\varepsilon,$

$b\varepsilon$) with a and b real numbers: if $|b| > |a|$, Q lies in T; if $a = b$, P lies on ℓ or ℓ'; if $|a| < |b|$, P lies in S. If we write $d = |a^2 - b^2|^{1/2}$, then in the first case, the infinitesimal distance between P and Q is εd, in the second, it is 0, and in the third it is $i\varepsilon d$.

Minkowski introduced "ict" to replace the "t" coordinate so as to make the metric of relativistic spacetime positive definite. This was purely a matter of formal convenience, and was later rejected by (general) relativists (see, for example Box 2.1, *Farewell to* "ict", of Misner, Thorne and Wheeler *Gravitation* [1973]). In conventional physics one never works with nilpotent quantities so it is always possible to replace formal imaginaries by their (negative) squares. But spacetime theory in SIA *forces* one to use imaginary units, since, infinitesimally, one can't "square oneself out of trouble". This being the case, it would seem that, infinitesimally, Wheeler *et al.'s* dictum needs to be replaced by

Vale "i$c(t)$", *ave* "iε" !

References

Bell, John L. *A Primer of Infinitesimal Analysis.* Cambridge University Press, 1998.

Misner, C.W., Thorne, K.S., and Wheeler, J.A. *Gravitation.* Freeman, 1973.

Moerdijk, I., and Reyes, G.E. *Models for Smooth Infinitesimal Analysis.* Springer-Verlag, 1991.