QUALITATIVE EXPERT EVALUATION AND QUANTITATIVE CHARACTERIZATION OF OFFICIAL REPORTS ON ALLEDGED UNIDENTIFIED AERIAL PHENOMENA IN FRANCE (1970-1979)

Jean-Pierre ROSPARS

Paper presented at the International Workshop CAIPAN
« *Collecte et Analyse des Informations sur les Phénomènes Aérospatiaux Non identifiés* »
organized by GEIPAN
Centre National d'Etudes Spatiales (CNES), Paris, France
8-9 July 2014

QUALITATIVE EXPERT EVALUATION AND QUANTITATIVE CHARACTERIZATION OF OFFICIAL REPORTS ON ALLEDGED UNIDENTIFIED AERIAL PHENOMENA IN FRANCE (1970-1979)

Jean-Pierre ROSPARS

Directeur de recherche
Jean-Pierre.Rospars@versailles.inra.fr

Since 1947, the existence of unidentified flying objects has been the subject of an intense controversy. Up to now, the only undisputable fact is that sighting reports on aerial phenomena considered as unidentified (UAP) by the witnesses, and sometimes by the investigators who interviewed them, are produced. In an attempt to evaluate the scientific potential of this production we have begun the examination of French sighting reports, to provide a synthetic description of their content and to attempt their characterization. The data analyzed come from the whole set of about one thousand reports issued by an official source, *Gendarmerie nationale* (French military division for law enforcement), on its own initiative or following direct witness notification, during the 1970-1979 decade, and collected in GEIPAN archives at CNES, Toulouse.

Our analyses aim at answering two linked questions:

- Does this set contain "interesting" sightings concerning phenomena that might be little or not understood (the signal) and is it possible to extract them reliably from the noisy background of "uninteresting" sightings concerning phenomena poorly observed or wrongly interpreted?
- Is it possible, once this qualitative screening is achieved, to find statistically significant quantitative differences between "interesting" and "uninteresting" sightings?

1. Reliability of report classification by "experts"

Can different experts achieve judgments of sufficient agreement so that we can trust their conclusions? We addressed this question in two steps:

First, we examined the expert evaluations performed in 1978 and 1979 by 28 engineers from CNES Toulouse on the reports of the period considered (1970-1979). Each expert has classified the reports he/she examined in four categories: (A) "Fully identified phenomenon", (B) "Phenomenon likely assignable to a known phenomenon", (C) "Unidentified phenomenon but the report is of little interest", (D) "Unidentified phenomenon and report of sufficient interest to deserve a subsequent analysis". Basically, D reports were considered unexplained despite the quality and quantity of available information.

Second, we achieved a personal evaluation of the whole set of reports using a similar classification with five categories – the same four categories as defined above, plus a fifth one "Weakly unidentified phenomenon", denoted C+, for reports of usually distant phenomena with a relative lack of descriptive details. Phenomena reminiscent of ball lightning (Piccoli, 2012) were classified as C+ or D.

These classifications denoted *G* (for GEPAN) and *J* respectively, are simple, easy and fast to implement, but the question is: are they reliable? A common *a priori* criticism insists on

their lack of formalization that gives the experts too much freedom. Evaluations, so goes the argument, will differ because reports have to be judged according to independent criteria, notably amount of available information, strangeness of the alleged phenomenon, and reliability of the witnesses, that are typically difficult to rank. Moreover, the final judgment will depend also on the ingenuity and biases of the experts. Therefore, doubts naturally arise on the validity of evaluations that seem to suffer from an over-dependence on "personal equations" of the experts. One may wonder whether similar cases are consistently put in the same category and different cases in different categories.

However, the reliability question can be approached in a more pragmatic way. A classification will be considered as reliable *if and only if the same report evaluated by two (or more) experts is put in the same category*. With this simple criterion the degree of internal consistency of the *G* evaluations and their consistency with the *J* evaluations can be determined and the degree of confidence that can be expected from expert evaluations can be estimated.

Fig. 1A shows that 77% of reports were evaluated by a least two *G* experts. For 65% of these reports with multiple evaluations, *G* experts were in agreement (Fig. 1B). If the two first "surely" and "probably explained" categories are put together the percentage of agreement rises to 81% (Fig. 1C).

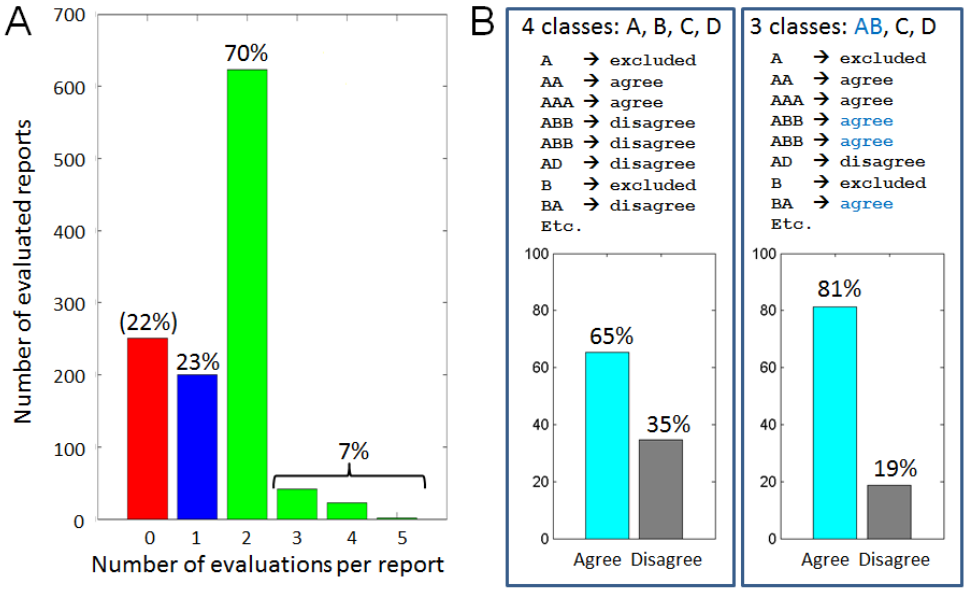


Fig. 1. Report evaluations *G* by CNES-GEPAN "experts". **A**. Number of evaluations by different experts. **B**. Proportion of agreements/disagreements for the reports with multiple evaluations with 4 (left) or 3 (right) categories.

The conflicting evaluations can be resolved by a conservative approach giving preference to the most critical judgments, so in the order D < A < B < C; for example if an expert considers a report as "unusable" (C), his/her conclusion will be preferred to that of another expert judging it as "probably explained" (B). With this reduction the two classifications *G* and *J* can be compared (Fig. 2), provided the *J* categories C and C+ are pooled together. Then, it is found that the global consistency is good (Fig. 2A); the difference for the "explained" reports is partly an artefact due to the events with multiple reports (see

"Technical notes" at the end); the difference for the "unexplained" reports (21% for *G*, 16% for *J*) is explainable by the "protective" effect of the C+ category (17% of reports). The detailed consistency based on an individual comparison of the *G* and *J* evaluations (Fig. 2B) shows that it is the same as that found between *G* experts for four categories, but slightly lower (68% instead of 81%) for 3 categories (with A and B grouped together).

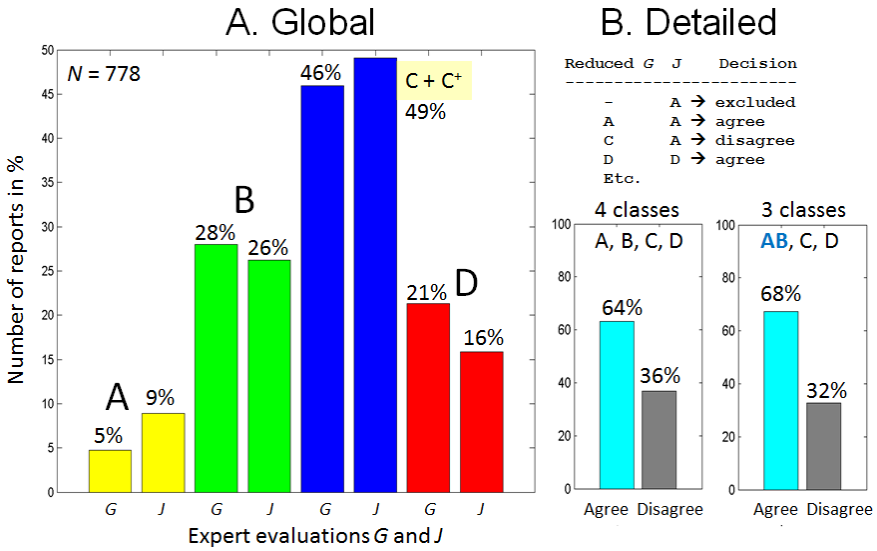


Fig. 2. Comparison of the reduced 4-category *G* evaluations with the 5-category *J* evaluations. **A**. Global comparison with my categories C and C+ grouped together. **B**. Detailed comparison with 4 and 3 categories.

In conclusion, the degree of agreement between experts is ~65% with four categories (A, B, C, D) and ~70% with the three basic ones (AB, C and D). A closer examination of data shows that contradiction between experts does not depend much on the category.

For future evaluations, it can be noted that the addition of category C+ seems to simplify the task of experts by decreasing their hesitations. The experts should be encouraged at mentioning their hesitation between two categories. The results above suggest that a relatively modest effort to make the evaluation procedure more explicit could improve its reliability.

2. Preliminary statistical analysis of reports based on expert evaluations

The aim of this second part is threefold:

- Put to use the evaluations analyzed in the first part and emphasize their importance for any scientific analysis of available reports;
- Provide a global description of a few basic aspects of the reported phenomena;
- Search for possible "signatures" distinguishing the identified A-B reports (control cases) from the unidentified D (and possibly C+) reports (test cases). The idea here is that all reports having been collected in similar conditions from similar witnesses, A-B reports can be used as control cases for possible differences with the D-C+ reports. A commonly held view is that the unexplained reports are undistinguishable from explained ones, suggesting they have in fact the same origin.

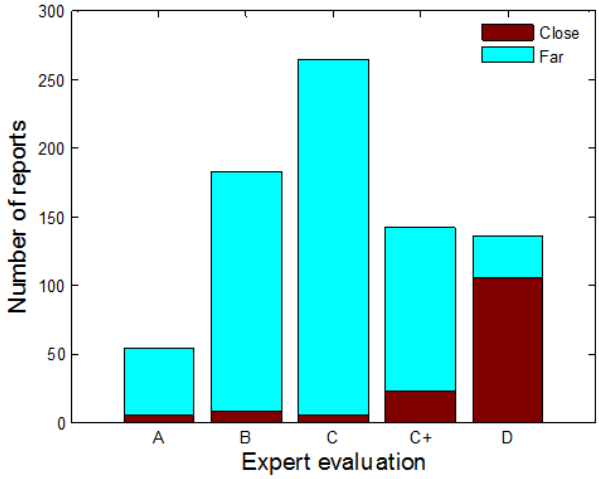
In the present preliminary study, only three basic descriptors will be considered – distance, location and time of the phenomena. The *J* evaluations with five categories: unusable (C),

surely (A) or probably (B) explained, more or less unexplained (D, C+) were utilized (*G* evaluations lead to similar results).

2.1. Distance

Is the identified/unidentified feature correlated to the distance between the observer and the reported phenomenon? We have distinguished events with unknown distance – mostly aerial phenomena – and with known or knowable distance – mostly phenomena on the ground or near the ground that were reported in the frame of reference of the witness, sometimes with corroborative evidence (phenomenon masking part of the landscape or illuminating a local spot etc.). For the sake of simplicity they can be dubbed respectively as 'far' and 'close'.

Most reported phenomena in the studied sample are 'far' (80%). More importantly, 95% of these 'far' phenomena are in categories A, B, C and C+, whereas 72% of the 'close' ones are in category D. This pattern is highly significant (Fig. 3). So, the distance between the observer and the alleged phenomenon appears as the major single feature that determined the expert classification.




```
Number of reports
 Far   = 95% ABC
 Close = 72% D
```

```
Number of reports
      Far Close
  ABC 602   41
  D    30  106
```

Fig. 3. Distance to the observers of the reported phenomena. In categories A, B, C and C+, the phenomenon is usually seen in the sky at an unknown distance ('far'). In category D, the distance is generally known or knowable ('close'). This pattern is highly significant (for the 2 x 2 contingency table shown based on n = 779 reports, X^2_{obs} = 372, p-value = 0).

2.2. Spatial distribution

How are the reported events related to the population of potential observers? In an earlier work (Rospars and Delécolle, 1978) based on a catalog of 400 American "close encounters" (Clark and Vallée, 1971) we found that, with respect to the population density of the American states, the number of events per square kilometre *increases* whereas the number of events per inhabitant *decreases*. We have reexamined the population-dependence for the various categories A-D at two spatial scales (regional and local).

First, we plotted the above densities of reports *N/S* (Fig. 4A) and *N/P* (Fig. 4D) for the 95 metropolitan *départements* against the density of population *P/S*, where *N* is the number of reports, *S* the area and *P* the population of each *département*. These log-log plots confirm the

population-dependence found previously for the American events. Therefore, the probability that an observation is reported in a (relatively large) area increases with the population of this area whereas the probability for an inhabitant to report an observation decreases when the density of the area increases. These two trends are characterized by the slopes of the regression lines in log-log plots, respectively 0.54 and −0.46. Now, are these slopes different for explained and unexplained reports? Figs. 4BC and 4EF show that the slopes of the regression lines of the four main categories (AB, C, C+ and D) are *not* statistically different. Therefore, the categories are undistinguishable as far as their dependence on the population of the *département* is concerned.

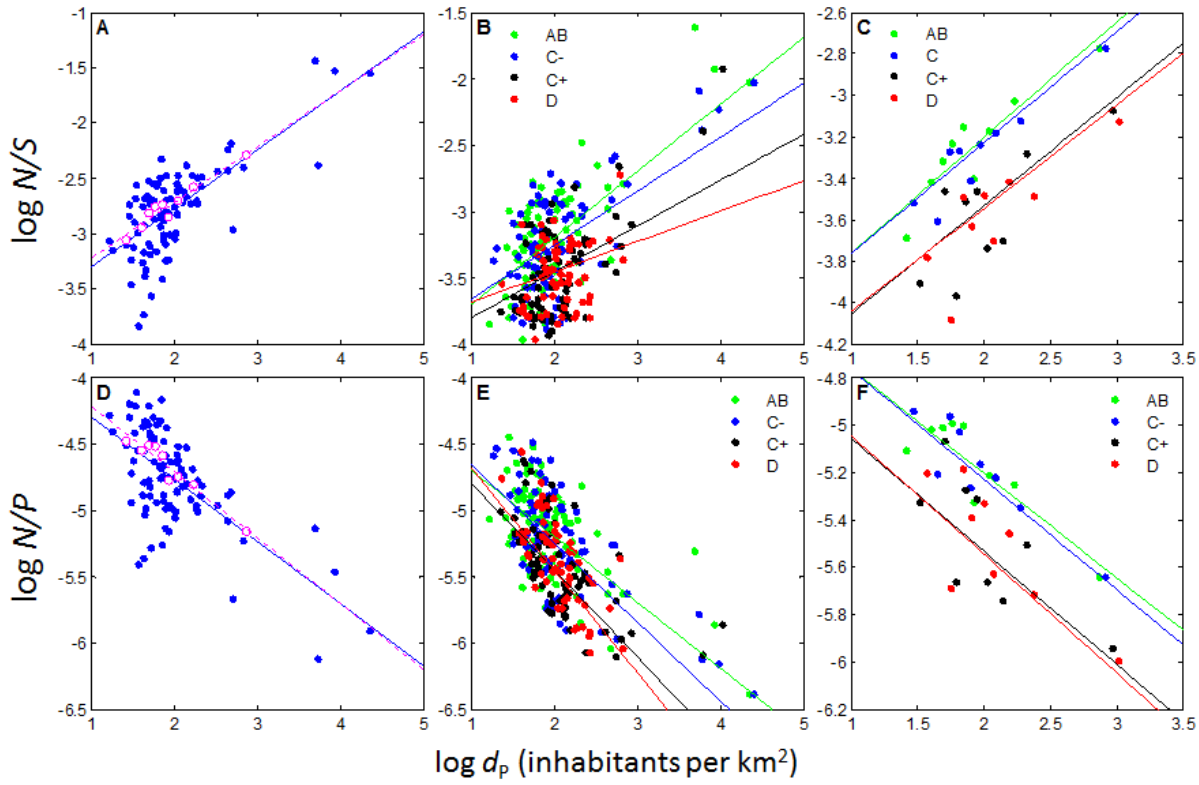


Fig. 4. Number of reports depending on the density of population of metropolitan *départements*. In all plots regression lines were determined after removing the 4 most densely populated *départements* (with log d_P > 3.3, that is d_P > 2000 inhab./km^2). **A**. Log-log plot of the total number of reports per km^2 versus population density d_P of each *département* (n = 94, blue points and solid regression line; one *département* has no report) and of *départements* with similar d_P's grouped in 9 classes (magenta circles and dotted regression line); in both cases slope ≈ 0.5, so $N/S \approx k\sqrt{d_P}$. **B**. Same as A for n = 94 *départements*, broken in 4 J categories (A and B grouped together); slopes of all regression lines are equal (Student's t tests, p-values > 0.06). **C**. Same as B for n = 9 classes of *départements*; slopes for AB (0.56) and D (0.50) are equal (t test, p-value > 0.7). **D**. Log-log plot of the total number of reports per inhabitant versus d_P, for n = 94 *départements* (in blue) and n = 9 classes (in magenta); in both cases slope ≈ −0.5, so $N/P \approx k'/\sqrt{d_P}$. **E**. Same as D for n = 94 broken in 4 categories, all slopes are also equal (t-test, p-value > 0.06). **F**. Same as E for n = 9; slopes for AB (−0.44) and D (−0.50) are equal (t test, p-value > 0.7).

Second, we examined the cumulated distribution of the number of reports with respect to the density of population d_P. It gives, for any density d_P, the cumulated number of reports occurring in all areas with a density smaller than d_P. This provides a very sensitive tool to uncover small differences between distributions when present. No difference was found between the explained AB and unexplained D cumulated distributions based on the density of population of the *départements* (Fig. 5A). However when the cumulated numbers are determined for the density of population of the metropolitan *communes* (the smallest administrative area in France), small but highly significant differences were found (Fig. 5B). This result suggests that the reports of category D occur preferentially in the least densely populated communes. This population-dependence appears at sufficiently high spatial resolution, of the order of a few km – the average linear dimension of a *commune* is 4.4 km – and disappears at coarser resolutions – the average dimension of a *département* is 80 km.

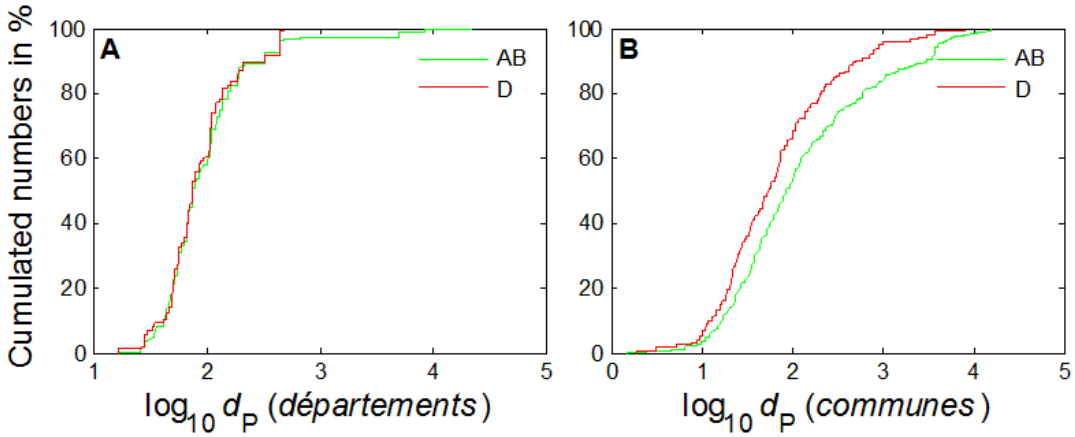


Fig. 5. Cumulated number of reports as a function of the population density. **A**. Distributions of explained (A, B) and unexplained (D) reports with respect to the population density of the 95 metropolitan *départements*. The two distributions are identical (Kolmogorov-Smirnov's test, *p*-value = 0.38) although they diverge for d_P > 477 inhabitants/km^2. **B**. Same distributions with respect to the population density of the ~36 500 *communes*; the two distributions are different (same test, *p*-value < 0.01).

Third, this effect encouraged us to reanalyze the data with respect to the density of population of *communes*. The communes with similar densities *P/S* were grouped together to form 100 classes. The ratios *N/S* and *N/P* in these classes and the corresponding regression lines are shown for all categories together (Fig. 6A, D), for each category separately (Fig. 6B, E) and for the "explained" AB versus "unexplained" D categories only (Fig. 6C, F). In the plot *N/S* vs. *P/S* (Fig. 6C) the slope for AB (0.60) is almost twice larger than for D (0.33) and their difference is highly significant. It suggests that, at fine spatial resolution, the number of explained reports per km^2 grows much faster with the population density than the unexplained ones. Similarly (Fig. 6F), the number of unexplained reports per inhabitant declines faster (slope −0.67) than the number of explained ones (−0.40).

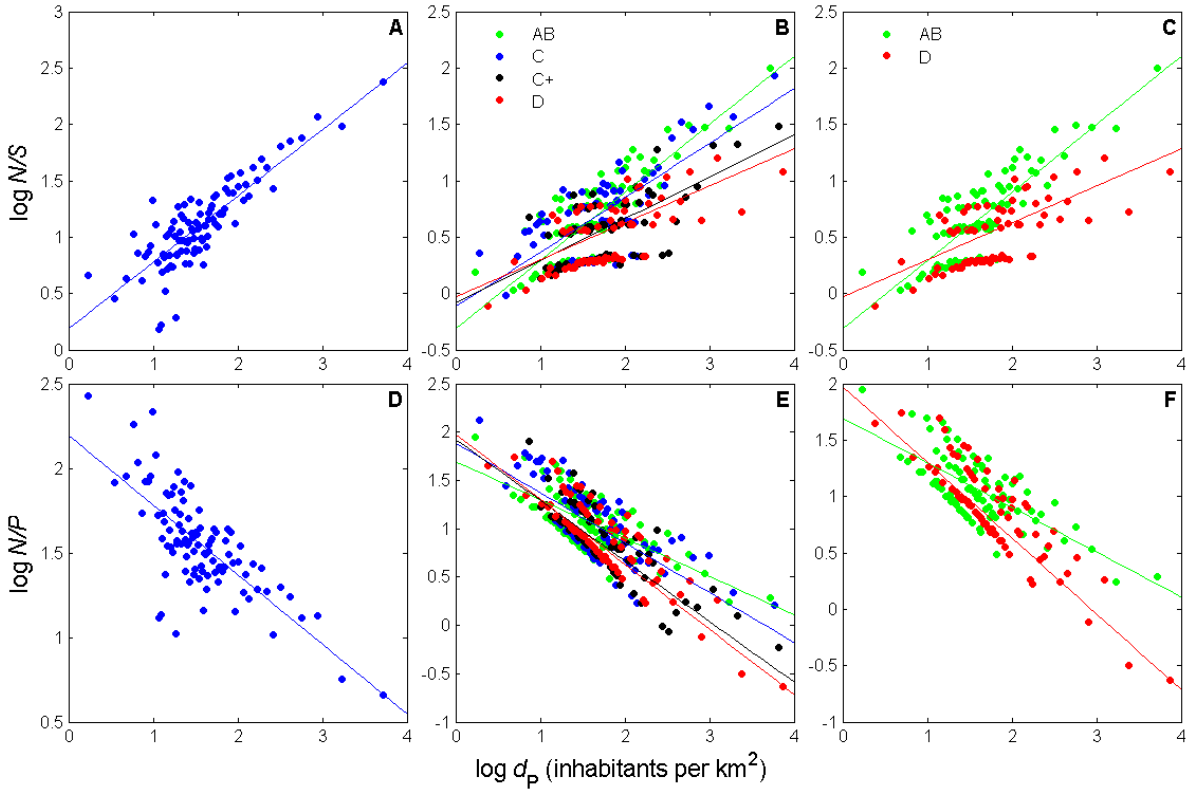


Fig. 6. Number of reports depending on the density of population of metropolitan *communes*. The ~36 000 *communes* with a similar population density d_P were grouped in 100 classes. Same representation as in Fig. 4. In C and F the difference between the slopes for explained (AB) and unexplained (D) reports is highly significant (Student's *t* test, *p*-value < 10^{-5}).

In conclusion, the slope of the regression lines *N/S* or *N/P* with respect to *P/S*, at fine (*communes*) but not coarse (*département*) spatial resolutions, appears to differ between the reports considered as unexplained by the "experts" and those considered as "explained".

2.3. Time distribution

How does the frequency of reports vary as a function of the time of the day?

First, the *legal time* distribution for all reports was established (Fig. 7). It shows a minimum around noon and three peaks, the largest in the evening from 9 to 10 p.m., the second in the morning from 6 to 7 a.m., and the smallest in the middle of the night from 2 to 3 a.m. This distributions is very similar to that found by Vallée in three different samples of "close encounter" events (Poher and Vallée, 1975). It is also very similar to the *universal time*, except that the UT distribution is one hour ahead of the legal distribution.

Second, in order to test this suggestion we separately plotted the legal time distributions of the five categories. A chi-square test based on the corresponding 5 x 11 contingency table (several hours had to be pooled in order to have all expected frequencies larger than 5) gave a large chi square value (793). Utilizing universal time instead of the legal time, modified the distributions but the chi square value varied little (731). So, the time distributions of the five categories, for the legal and universal times, are different from those expected under the null hypothesis of identical distributions.

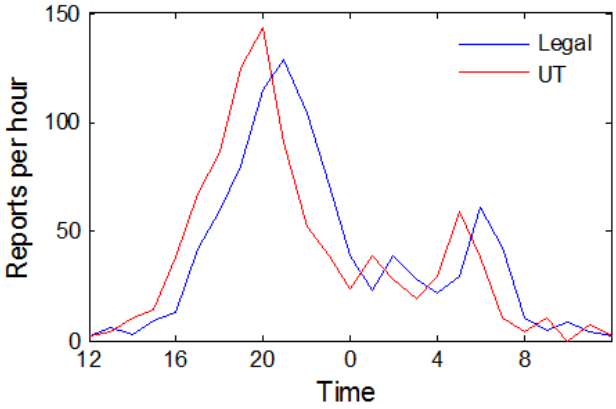


Fig. 7. Time distribution of the reported phenomena. Legal time (n = 950) and universal time (n = 938), for all reports. The number of reports with an observation occurring, for example, between 21h00 and 21h59 is plotted at time 21h00 (9 p.m.).

Third, the time distributions of the explained AB and unexplained D reports were compared. Again, the chi squares are large, for both legal (592) and universal (664) times. Fig. 8 shows the difference between the two distributions; the expression in percent (Fig. 8B, D) shows a clear excess of D cases over AB cases from 9 p.m. to 3 a.m. compensated by a deficit at most other times.

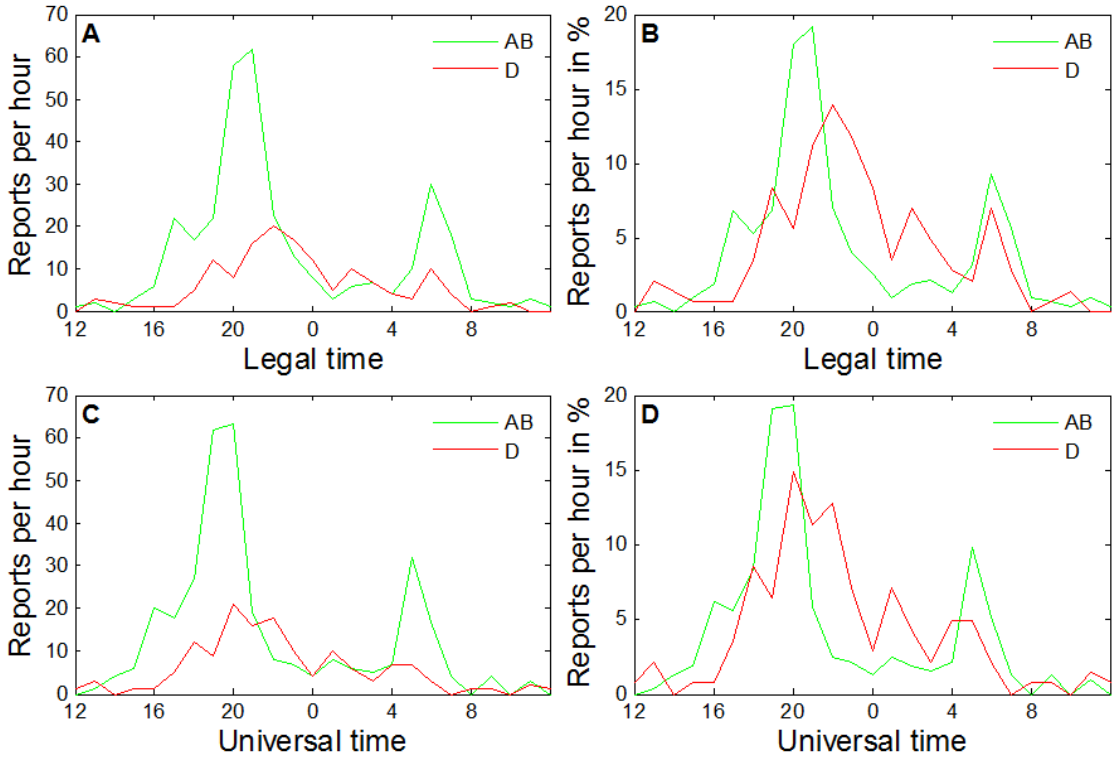


Fig. 8. Comparison of the time distributions of the "explained" an "unexplained" phenomena. **A**, **B**: legal time. **C, D**: UT. For both times, the differences between the 2 distributions are highly significant (2 x 13 contingency tables based on n = 357 reports, X^2_{obs} = 592 for legal time and X^2_{obs} = 664 for UT, p-value = 0).

Future investigations

In conclusion, the present study suggests that a preliminary evaluation of reports may (i) be approved by various specialists and (ii) be confirmed *a posteriori* by statistical analyses of parameters (place and time of observations) independent of those utilized to evaluate the reports. However, it leaves aside many aspects and raises many questions. Part of them concerns the methods of evaluation:

- The results at hand encourage to formalize these methods. For example they call for a better definition of the "surely explained" A and "probably explained" B categories. We have considered that an explanation is sure either when it is based on verifiable arguments (typically when the phenomenon can be clearly identified as the moon, a planet or a bright star) or on facts found by field investigators, and "probable" when it relies on assumptions.
- They suggest that the "weakly unidentified phenomenon" C+ category might be useful because it helps the expert to have a more consistent "unusable" C category of reports and because C+ reports might present the same trends as the D reports.
- Reproducibility of expert evaluations for different judges has been tested but not their repeatability (does the same judge produce always the same judgment?). This problem arises because evaluations are usually extended over a long period of time (especially *J* evaluations).

Other questions concern the observed phenomena:

- Are "explained" and "unexplained" reports really indiscernible (Hendry, 1979)? The preliminary results reported here do not confirm this widely-held statement. May methodological problems related to sampling, choice of criteria or statistical tests explain these differing conclusions?
- The analysis of report emission in the framework of the potential witness approach seems promising. Pioneered by Vallée, it led him to suggest that "the decrease in reports of close encounters between 11 p.m. and 2 a.m. may simply be due to the fact that the number of potential observers falls drastically as most people spend these hours at home" (Poher and Vallée, 1975). Are A and B reports following the use of time established by sociologists? Can they be used as an internal control to estimate the actual probability of occurrence of D phenomena?
- More generally, may the combination of expert evaluation and statistical comparisons lead to a more objective appraisal of the global significance of reports?

All these questions (and others) are better left to future investigations based on more extensive databases and more precise models taking into account more parameters. At least their development does not seem the waste of time that was feared when this study was begun and that, for so long, discouraged studies of this kind.

Technical notes

This preliminary study is based on the whole set of 983 *procès-verbaux* (official reports) produced in the 70's by *Gendarmerie nationale* and available at GEIPAN. The earliest event occurred on 20 February 1970 and the latest on 29 December 1979. All reports were read and

coded by the author. For the sake of simplicity, when several events were reported in a *procès-verbal*, a single one was kept for further analysis. When the same sighting (same witnesses) generated several reports, a single evaluation was done. When several reports were generated by a single event (mostly meteor or satellite re-entry), multiple identical per-report evaluations (*J*) and, in general, unique per-event evaluations (*G*) were done. A few reports (less than 2% for *J*) were not evaluated (ground traces without sighting, illegible copies, unfinished analyses). A more consistent treatment of witnesses, events and reports will be attempted in future versions of this work.

Administrative and demographic data were obtained from INSEE (*Institut National de la Statistique et des Études Économiques*). They are based on the 1975 census. Several population counts per *commune* being provided, the "total legal population with double counts" was used (the results are practically the same with the other counts).

Data were processed with Matlab and the Statistics Toolbox (The MathWorks, Natick, USA), supplemented with several custom libraries (data management, statistics, time, graphics etc.)

Acknowledgements

I thank C. Poher, A. Esterle, J.-J. Vélasco, J. Patenet and X. Passot for giving me access to the GEPAN/SEPRA/GEIPAN archives and for helpful discussions. I thank also J. Pagès, R. Delécolle and J. Vallée for critical comments on the manuscript.

References

Clark, J. and Vallée, J. (1971) Researching the American landings. *F.S.R.* 17, n° 5 and 6.

Hendry, A. (1979) *The UFO handbook. A guide to investigating, evaluating and reporting UFO sightings*. Doubleday, New York.

Piccoli, R. (2012) A statistical study of ball lightning events observed in France between 1994 and 2011. Lightning Strike Research Laboratory, 15270 Champs-sur-Tarentaine, France. http://www.labofoudre.com, 6 pp.

Poher, C. and Vallée, J. (1975) Basic pattern in UFO observations. AIAA paper 75-42, *13th Aerospace Sciences Meeting*, Pasadena, 14 pp.

Rospars, J.-P. and Delécolle, R. (1977) Recherche de modèles de répartition dans l'espace et dans le temps d'atterrissages allégués d'O.V.N.I. aux Etats-Unis. *Présentation au conseil scientifique, CNES-GEPAN*, Annexe 12, 31 pp.