Data Mining in the SIMBAD database web log files
Wenger M., Oberto A.
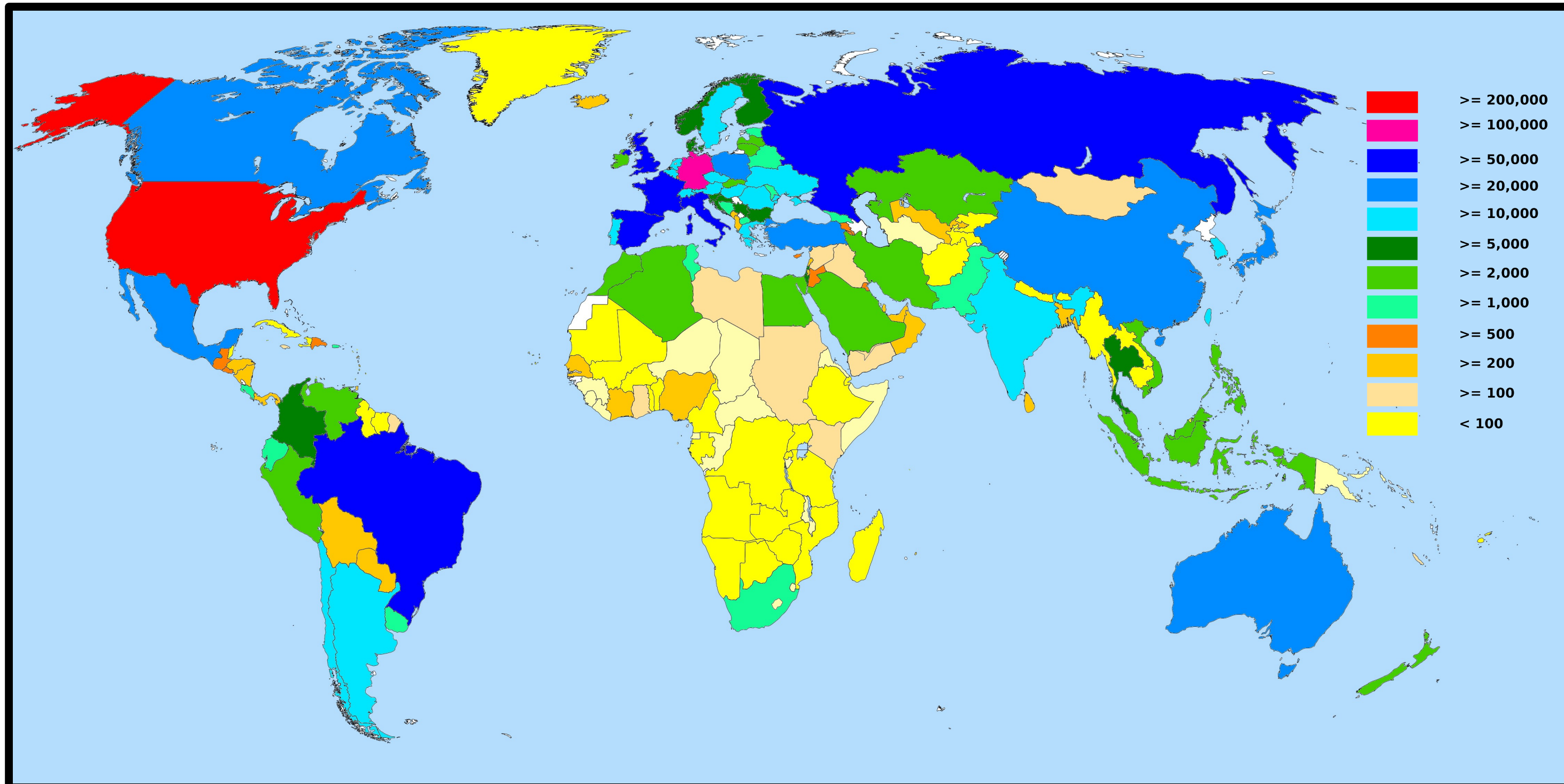## Centre de Données astronomique de Strasbourg

Introduction

The SIMBAD web server produces a very large amount of information in its log file. The available information consists of IP addresses, time of the query, query type and content, and information on the user's environment (hardware, operating system and browser).

This presentation is based on the SIMBAD web server log files collected between January 2007 and June 2009. It shows some information rarely extracted from such logs.

Users and countries

With the help of a geolocalisation database associating IP addresses with the corresponding geographic location (country, city, coordinates, timezone) [ref 1], it is possible to associate most of the IP addresses with a country.

With the approximation that one IP address correspond to one user, we can count the number of different users in each country. Over two and a half years, this leads to the following map:

| |
|---|
| >= 200,000 |
| >= 100,000 |
| >= 50,000 |
| >= 20,000 |
| >= 10,000 |
| >= 5,000 |
| >= 2,000 |
| >= 1,000 |
| >= 500 |
| >= 200 |
| >= 100 |
| < 100 |

223 countries, territories and islands have accessed SIMBAD in the last 30 months. Red and blue covered countries correspond probably to the most active countries in astronomy. Surprisingly, there are very few countries remaining in white (no user at all) on this map.

Users working environment

What kind of hardware, operating system and browser are used by the SIMBAD users ? The Apache log allows to answer these questions after having defined a few rules to infer unspecified information (i.e. MacIntosh hardware can be deduced from MacOS operating system). The figures are associated with unique IP addresses.
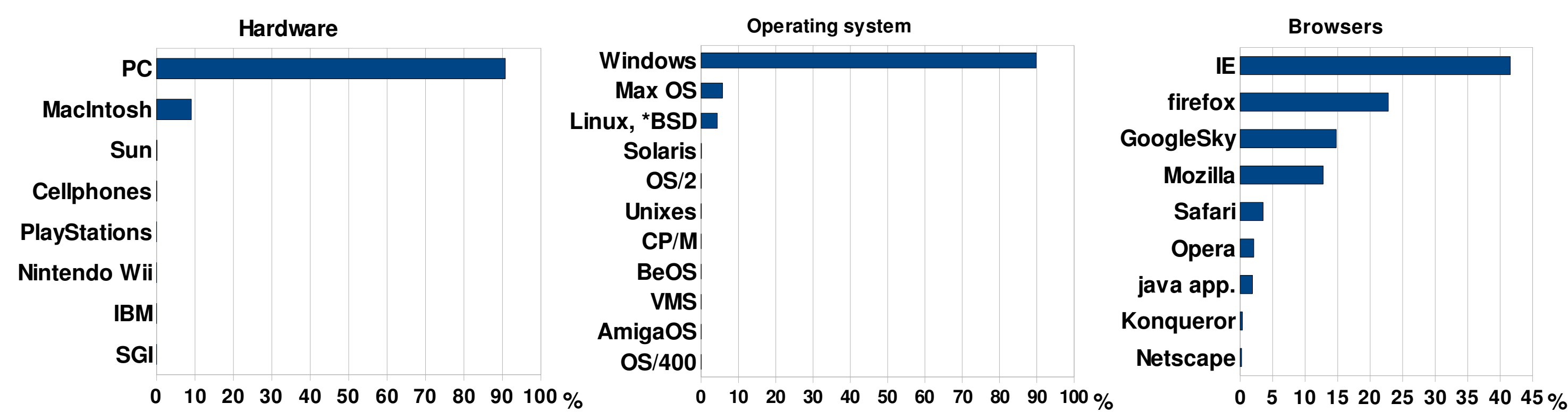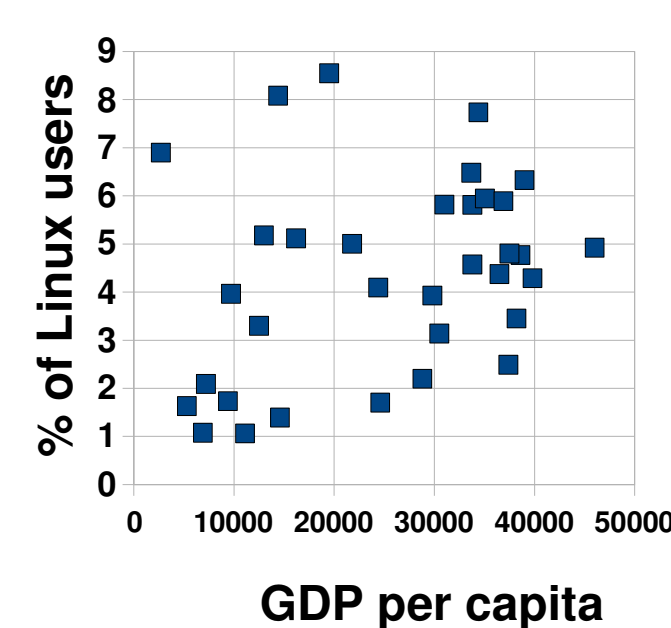

PCs and Windows are clearly dominant, but there are some clear differences among the countries as it is shown below for the 34 countries having the most users :
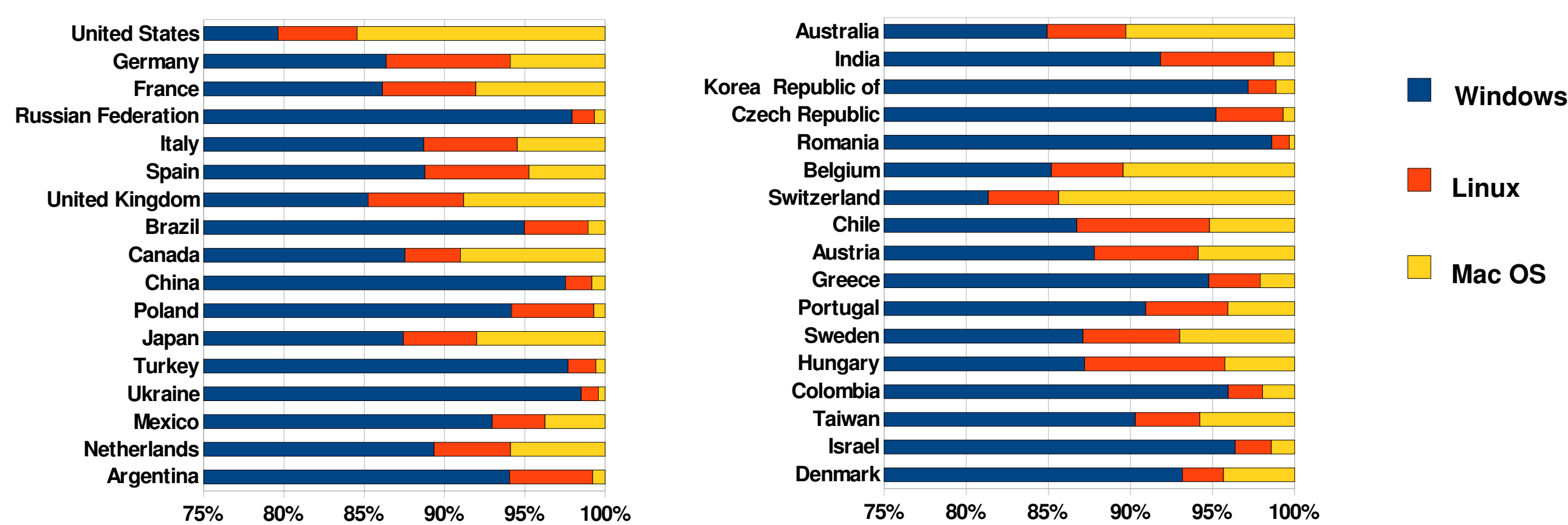
There is a correlation between the GDP (Gross Domestic Product) per capita [ref 2] and the percentage of Mac Os users in these countries
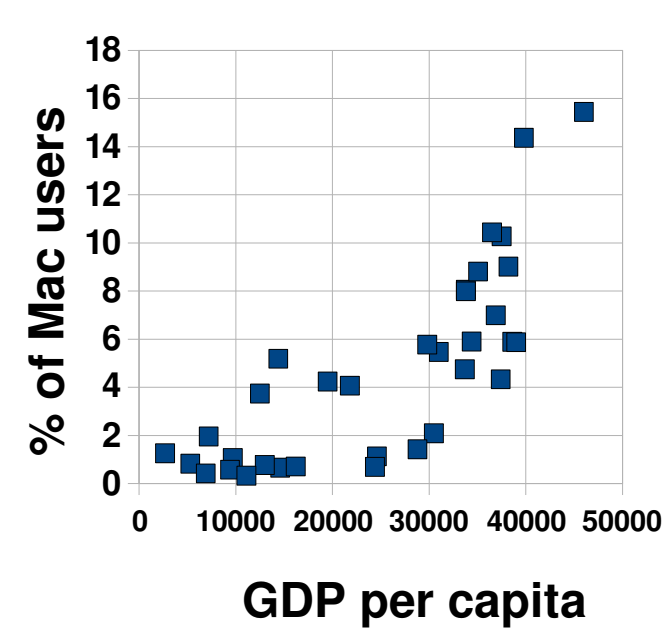

This correlation does not exist for Linux users...

Session duration

Web accesses are stateless. Nevertheless, it is possible to infer some information concerning the time a user is spending querying the database. This is done by checking all requests from one IP address and considering that a session is ended if the time between two requests is over a given threshhold, which is here defined as 30 minutes.
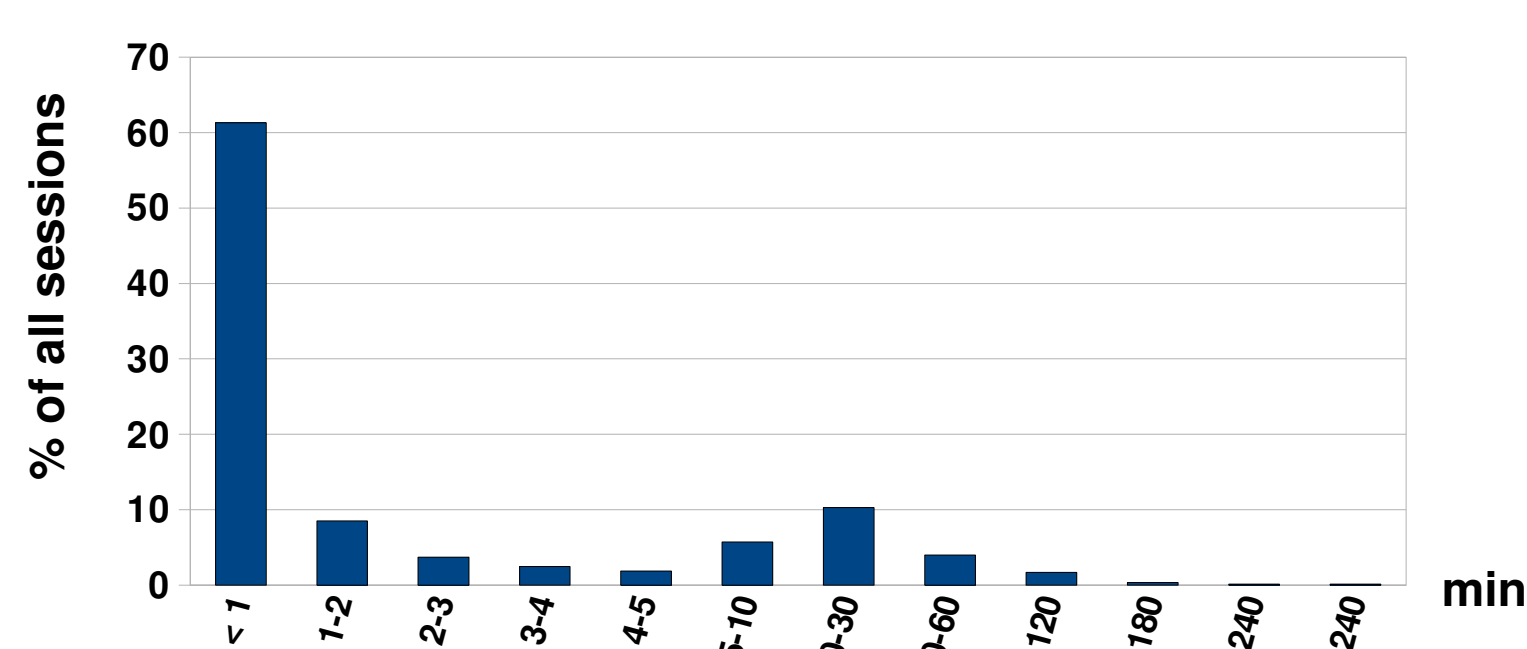

Session time

At which time of the day is SIMBAD the most used, or the less used ? Can we find out when users work in different countries ? Here are some answers.

Percentage of requests executed every hour (CDS local time) on workdays. The minimum is in the morning at 4-5 am. The maximum is at the begining of the afternoon, when also the users of the american continent begin to work. The decrease during the evening is slow.
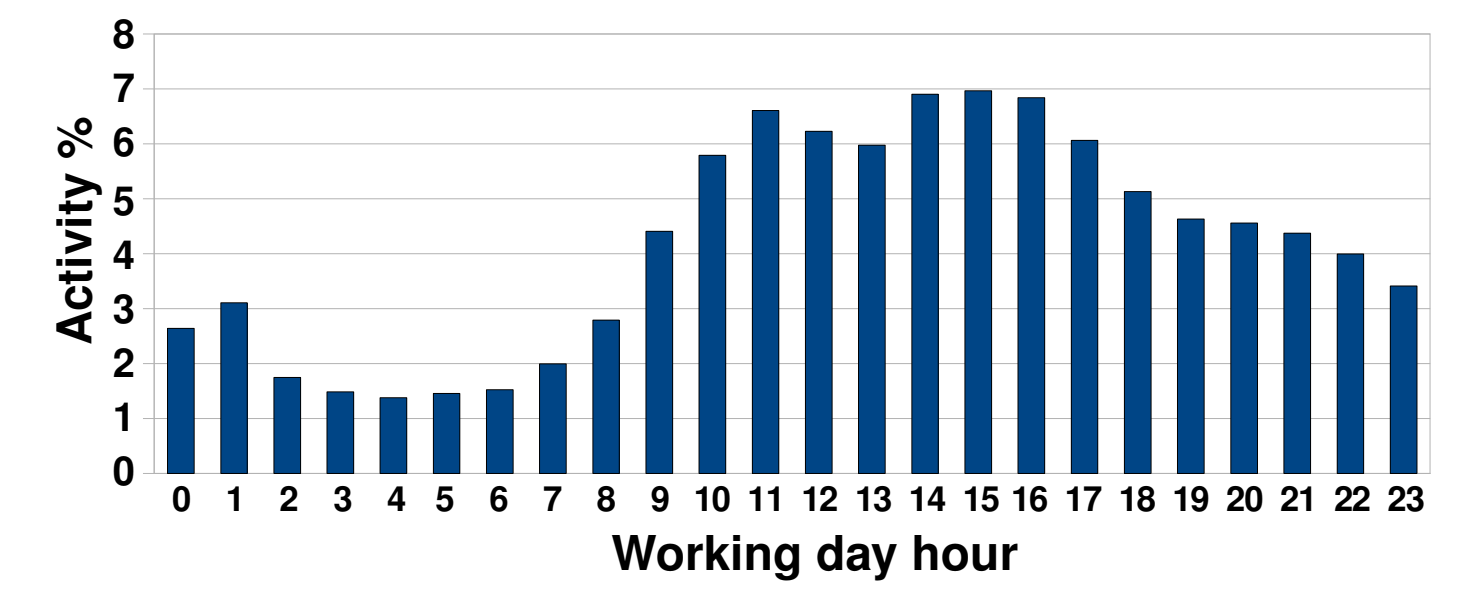

The IP address allows us to infer the country of each access. Taking also in account the rules defining for each country its Daylight Saving Time periods, we can analyze the working hours in different countries.
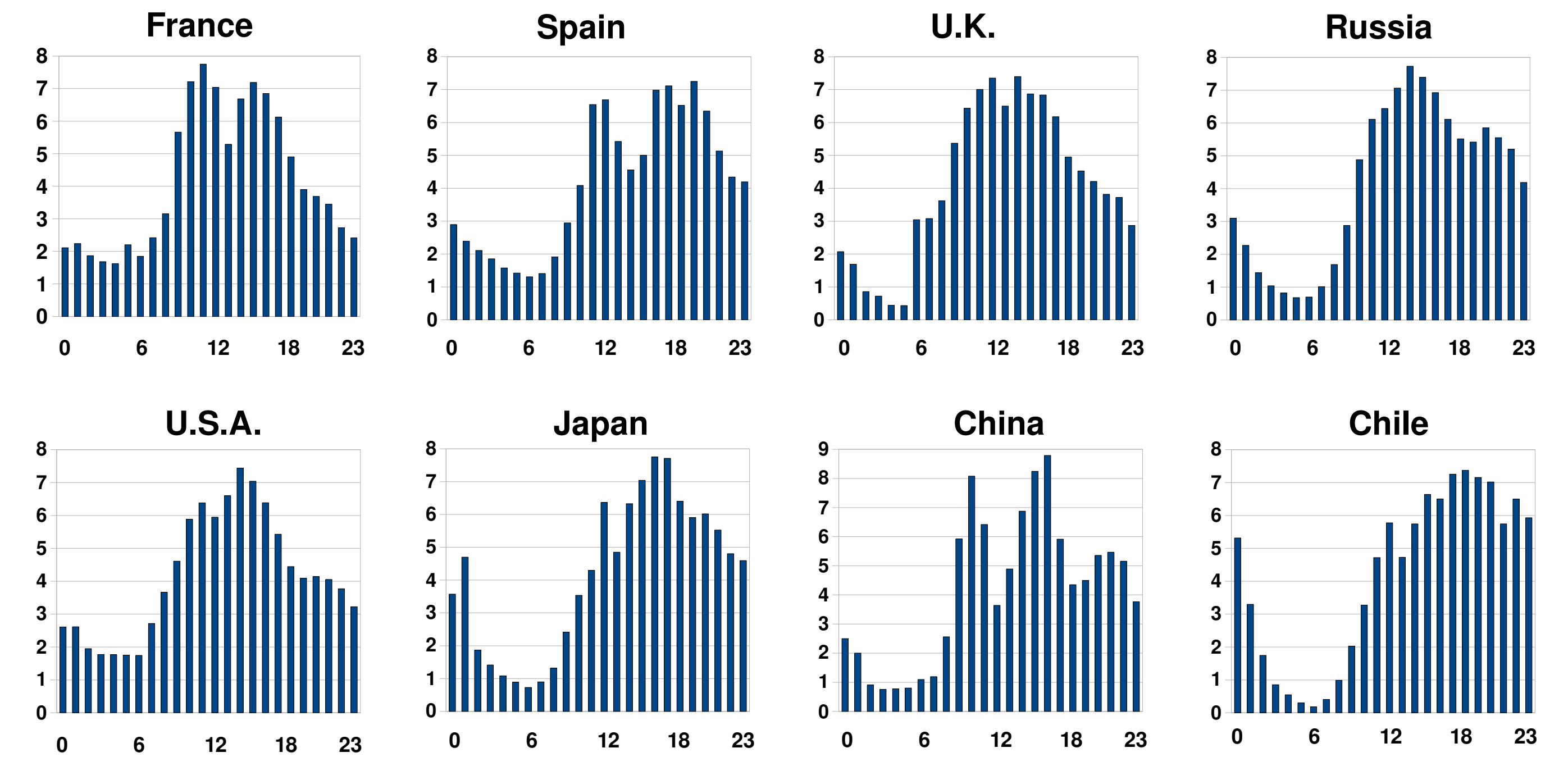

Queries

The Apache log contains also the query parameters from GET commands. Several information can be extracted from them, namely the identifiers and catalogs used for identifier queries, and the main object types associated with these catalogs.
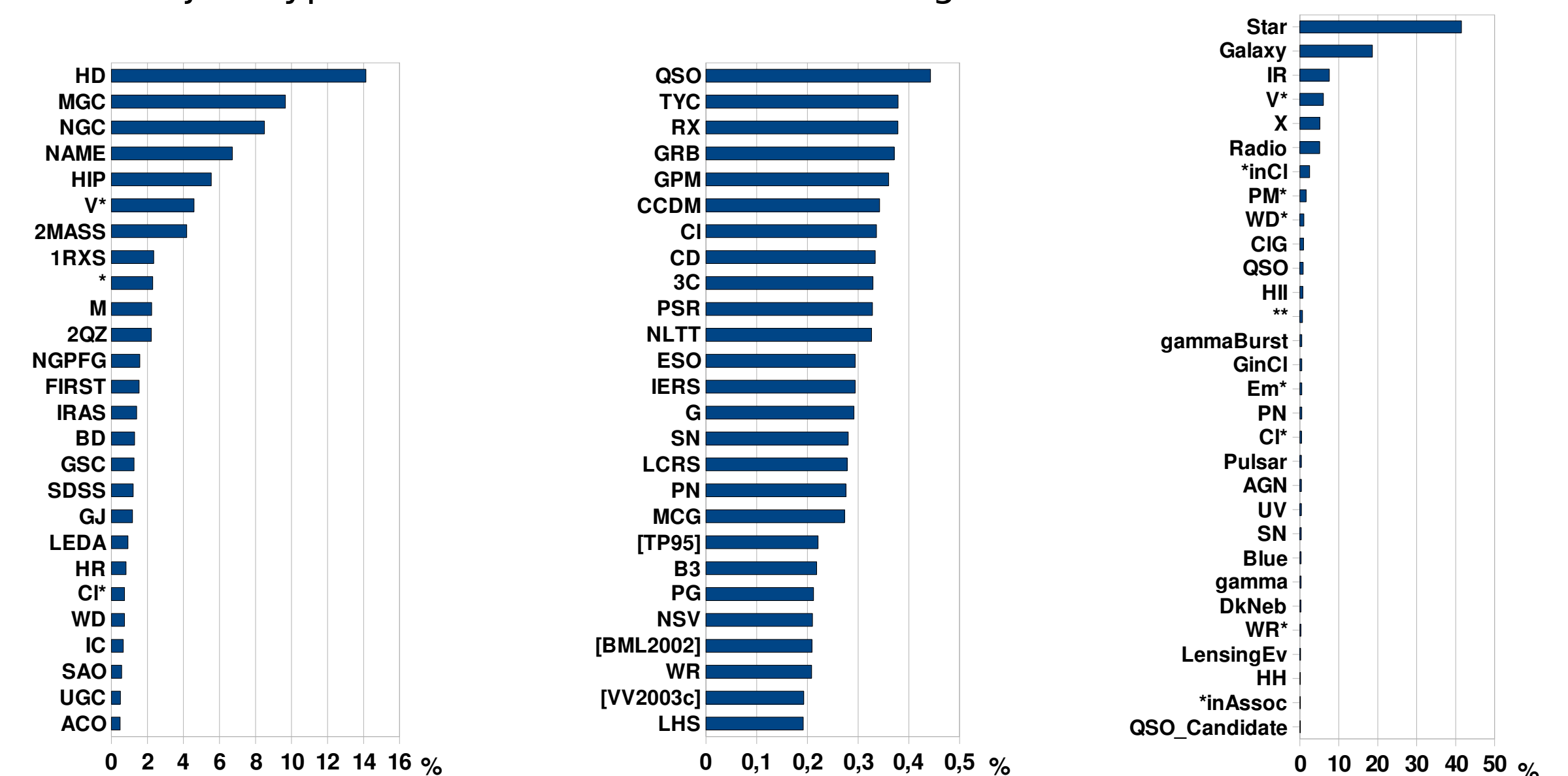
In two and a half year, about 1,500,000 different identifiers were queried. The most popular names are listed below

```
The ten most popular names :
NAME SMC
M 31
NAME SIRIUS
HD 8890 = POLARIS
M 33
NAME VEGA
M 101
HIP 32349 = SIRIUS
NAME LMC
NAME SGR DSPH
```

```
Object names :
LMC
SMC
ALCYONE
ALDEBARAN
ALGOL
ALTAIR
ANTARES
ARCTURUS
BETELGEUSE
CANOPUS
CAPELLA
CENTAURUS A
CRAB NEBULA
DENEB
EAGLE NEBULA
FOMALHAUT
HYADES
KOCHAB
MIRA
PEANUT NEBULA
PLEIADES
POLARIS
PROCYON
PROXIMA CENTAURI
REGULUS
RIGEL
SAGITTARIUS DWARF GALAXY
SCULPTOR FILAMENT
SGR A
SIRIUS
SPICA
VEGA
```

```
Various objects :
NGC 253
NGC 1068
NGC 2362
ACO 1689 (galaxy cluster)
NGC 2264

3C 273
X Cyg X-1
SS 433
Granat 1915+105
* alf Cen A
* alf Cen B
* eta Car
* alf Ori
V* alf Lyr
V* TW Hya
GJ 581 (extrasolar planets)
GJ 436
GJ 876
HD 209458 (= V* V376 Peg)
HD 121409
HD 125019
HD 189733 (= V* V452 Vul)
HD 217107
HD 148478 (= V* alf Sco)
```

```
Messier objects :
M 1       M 45
M 3       M 51
M 13      M 57
M 15      M 67
M 16      M 80
M 17      M 81
M 31      M 82
M 32      M 83
M 33      M 87
M 42      M 101
M 44      M 104
```

Queries by coordinates are defined by a center and a radius. These maps show the most queries areas in the sky :

Equatorial coordinates

Galactical coordinates
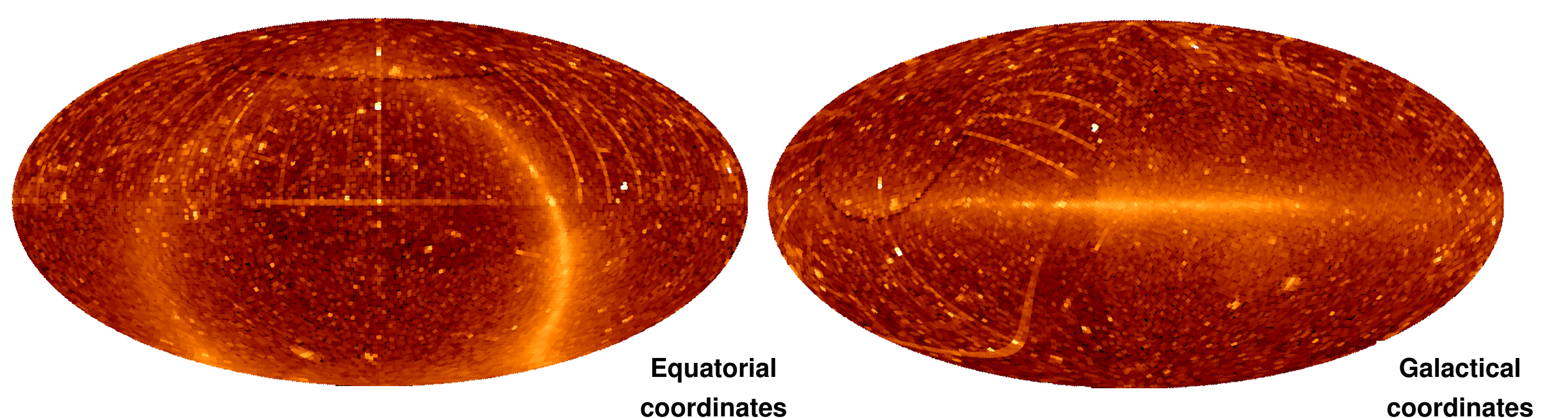
References

[ref 1] Geolocalisation database : http://ipinfodb.com
[ref 2] Gross Domestic Product per capita database: http://www.indexmundi.com/