Machine Learning From Archives Gauquelin

Abstract

We apply machine learning methods to the data from Archives Gauquelin in an attempt to build a binary classifier able to distinguish between outstanding scientists and sports champions using only astronomical factors derived from their natal data. We apply a special splitting into training, validation and testing sets, and a set of three combined features, each of which combines dozens of elementary astronomical features. Our null hypothesis is that accuracy on Testing sets must be 0.5 if the Training sets contain the same number of group A and group B representatives born on each year, that is, if yearly frequencies are equal. Our external Testing sets contain only persons born later than those in Archives Gauquelin. Logistic Regression is our primary method, and Random Forest an alternative. All data and implementations of our algorithms are available from a public repository on GitLab.com.

Introduction

Archives Gauquelin (AG) contain birth data of outstanding professionals, including Sports Champions (SC), Scientists And Medical Doctors (SMD), and other groups, for example, Mental Patients and The Hereditary Experiment Subjects. SMD is the biggest professional group.

Data were collected by Françoise and Michel Gauquelin in 1949-1991, they used the data to investigate and report their findings, the most known of which is the so-called Mars Effect: a statistically significant number of **sports champions** were born just after planet Mars rises or culminates, relative to the horizon at time and place of birth.

We use data from AG as published on the C.U.R.A. web site[1], but we do not use astronomical factors that depend upon the daily rotation of the Earth.

Trying to build a classifier for professionals with only astronomical factors as features is certainly a questionable direction. First of all, the common knowledge is that all prior research “failed to find effects commensurate with astrological claims”[2]. If classifier could label better than a random number generator, that would support the main astrological claim.

Also, why professionals? If there was any correlation between astronomical factors and something about behavior of humans who were not aware of those factors, then most likely the correlation would be stronger between astronomical factors and something psychological, like personality traits, or biological/physiological, like gene expression patterns or microbiota activity patterns.

Furthermore, lets assume there may be correlations between astronomical factors at the time of birth, and *something physiological/psychological*, and then correlations between the latter and professions in which people succeed. In this case, correlations with professions would probably vary significantly over time: in the 19th century a noticeably different set of personality traits would work best for becoming a sports champion, or an outstanding scientist, than the set of traits that works best in the 21st century.

We decided to try building a classifier using data from Archives Gauquelin because AG are easily available online, and have been available for many years, and because a number of relatively recent studies[4][5][6][7] suggest that there may be a lot more to learn about physical processes in the Solar System and their impact, especially about the imperceptible impact, on humans and other species. We understand that overall a genome that is able to utilize both A and B, two features of the environment, can become better adapted to environment than a genome able to utilize only A but not B, or vice versa, only B but not A.

Fetching data and converting them to MLAG format

Is all done completely automatically using the Python scripts we provide[8]. Plots are there as well.

Features

Our null hypothesis is as follows: classification accuracy on the Testing set must be close to 0.5 if every Training set contains the same number of group A and group B subjects born on each year. That is, if yearly frequencies are equal. We try to prevent classifier from learning the yearly frequencies using three different methods.

First, with a special splitting into Training, Validation, and Testing sets. In every Training set the yearly frequencies are equal.

Second, we were building a classifier able to distinguish between outstanding scientists and sports champions, but during the Parameter Optimization process we used seven other pairs of groups in addition to the target pair (Scientists And Medical Doctors, Sports Champions).

Third, we avoid using astronomical features that (supposedly) help classifier to learn the yearly frequencies: features that last for months or even years. That is, features with only Jupiter and planets beyond it, with no faster moving bodies: Moon, Sun and planets closer to Sun than Jupiter.

We also do not use astronomical factors that depend upon the daily rotation of the Earth (and thus depend on time of birth) including factors of the class “planets in twelve astrological houses”. This makes our research less dependent of the possible bias in AG, and this also makes our classifier more robust to unknown or imprecise time of birth.

Our three combined features correspond to the three nearest cosmic neighbors of Earth, namely Moon, Venus, and Mars. These are also the first three columns in the summary table with findings of Gauquelins and the researchers who tried to replicate and further investigate their findings [3].

Each of the three features is a weighted sum of elementary features from two classes of factors being used in Western astrology since ancient times: astrological aspects, and planets in twelve Zodiac signs (*our only* truly *innovative components are the secondary elements of celestial bodies, with aspect weights depending on both the primary and secondary elements*).

For example, the value of the “Mars” feature is a sum of a parameter corresponding to Mars’ position in one of the twelve 30-degree sectors (ecliptic longitude quantized to 12 values), and the addendums originating from special angles, so-called astrological aspects, between Mars and seven major celestial bodies, namely Sun, Moon, and planets up to Uranus, excluding Neptune and Pluto:

Sun, Moon, Mercury, Venus, Jupiter, Saturn, Uranus

The sets of bodies and aspects are the same for the other two features, “Venus” and “Moon” (with Mars in place of Venus or Moon in the list of bodies), but 28 other parameters are different.

The total number of parameters is 122, with 28*3 of them impacting just one of the three features, and the remaining 38 parameters having an impact on all three features. 30 of the 38 are the weights of aspects and *allowances* of aspects, that is, *half-widths of sectors*, also known as *orbs* in Western astrology: for example, “body is in an opposition with Mars, with a 4-degree orb” means the angle is anywhere between 176 and 184 degrees between (the ecliptic longitudes of) the body and Mars. In this example, width of sector is 8 degrees, and orb is 4 degrees.

Another example, the parameters corresponding to the Zodiac sign at which Moon was at the moment are as follows: [0, 3, 2, 1,-1, 0, 2, 2, 1,-1, 2,-8]. That is, +3*w to value of the “Moon” feature if the ecliptic longitude of Moon, expressed in degrees, was between 30 and 60, and -8*w if it was between 330 and 360, where w is the Moon’s Zodiac multiplier parameter, w=5.5. As can be seen from this example, some of the 122 parameters are redundant: here we could have 12 instead of 13 parameters.

Training, Validation and Testing sets

For each pair (group A, group B) we extract a Testing set using the following method. For every year Y, we look at the numbers of persons born on that year, NA[Y] and NB[Y]. If NA[Y]=0, then all of the persons from group B born in year Y are appended to the Testing set, and vice versa, if NB[Y]=0 then all NA[Y] persons from group A born in year Y are appended to the Testing set. The joint Training+Validation set is the other outcome of this procedure.

Then during the Parameter Optimization (PO) process, as well as after it is complete, during the Testing set evaluation process, every trial is as follows. We repeat N times: split the Training+Validation set into a Training set and a Validation set, using a simple method described in details below, then train a classifier with the Training set, and evaluate it on the Validation set, and report the Training set accuracy and the Validation set accuracy. Note that PO process takes weeks on a modern laptop using the Logistic Regression implementation from the Scikit-learn machine learning library[12].

We use N=3000 in every trial for each pair (group A, group B), because splitting into Training and Validation sets is done with a pseudo-random data generator: for every year Y, if NA[Y] > NB[Y] then we select NA[Y]-NB[Y] persons from group A at random, and push them into Validation set. If NB[Y] > NA[Y] then we select NB[Y]-NA[Y] persons from group B at random, and these are appended to the Validation set.

In either case, the remaining M*2 persons, where M = min(NA[Y], NB[Y]), are pushed into the Training set. Thus, the yearly frequencies in the Training set are equal.

The outcome of every trial is a pair of numbers: the average Training set accuracy, ATraSA, and the average Validation set accuracy, AVaSA. The median Validation set accuracy is also reported, but it is not used in the Parameter Optimization process.

The average Testing set accuracy ATeSA is also an arithmetic mean from 3000 iterations, calculated exactly like AVaSA: for each splitting into Training and Validation sets, evaluate on the Testing set, then take an arithmetic mean from 3000 iterations. Testing sets are ignored during the PO process.

Pairs of groups and the parameter optimization target

We built seven groups from AG data:

SC: Sports Champions
SMD: Scientists, Medical Doctors
Military Men
Merged 6: all actors, journalists, musicians, painters, politicians, writers from AG
Heredity Experiment subjects from Volume B
Heredity Experiment subjects from Volume E2
Mental Patients

During the PO process each PO iteration is as follows. Run the trial with N=3000 on each of the following **eight pairs,** and report the average of eight ATraSA values, plus the average of eight AVaSA values. The average of these 16 values represents the outcome with the given set of parameters. Thus the parameter optimization target is the average of 48000 values.

Group A	Group B	Training set size	Validation set size from group A	Validation set size from group B	Testing set size from group A	Testing set size from group B
SC	SMD	1414 x 2	897	1937	575	1364
SC	Military men	1374 x 2	971	1350	541	1195
SMD	Military men	3392 x 2	1310	525	*13**	*2**
SC	Merged 6	2329 x 2	547	4148	*10**	*2724**
SMD	Heredity vol. B	3383 x 2	452	17776	880	3150
SMD	Heredity vol. E2	2335 x 2	1048	24438	1332	760
SC	Mental patients	2310 x 2	331	2209	*245**	*1**
SMD	Mental patients	1622 x 2	1075	2721	2018	177

Table 1. Sizes of Training, Validation, Testing sets. *These Testing sets are never used.

AVaSA and ATraSA are summed up with equal weights, this represents our intention: at the next parameter improvement step we want a parameter set with a higher AVaSA, but only if it does not decrease ATraSA by a bigger amount: a decrease in ATraSA is allowed if it is smaller than increase in AVaSA. Similarly, a set with a higher ATraSA, but only if it does not decrease AVaSA even more.

Results

We continued PO process until the optimization target exceeded 55.2%. We were using only the Logistic Regression function from the Scikit-learn library with all the default parameters, except solver='liblinear': *classifier = LogisticRegression(solver='liblinear')*. With the obtained “final” set of parameters the eight ATraSA values are between 53.6% and 57.7%, on average **54.46**%. The eight AVaSA values are between 54.13% and 59.74%, on average **55.96**%. Seemingly the classifier has learned too many peculiarities of our Validation sets during PO, probably because they are built such that for each year Y, either group A or group B subjects born in year Y go to a Validation set.

The average Internal Testing Set accuracies, after 3000 iterations as always, are as follows:

57.92% of 575 sports champions were classified correctly as sports champions;

55.30% of 1364 scientists and medical doctors were classified correctly.

On the External Testing Set, that is, persons from SADC[9][10][11] born after persons from AG:

57.57% of 235 sports champions were classified correctly as sports champions;

58.36% of 51 scientists and medical doctors were classified correctly.

Another interesting result is **the signs of aspect weights**. Initially, all weights were zeros, and at every iteration of the Parameter Optimization process they could either increase or decrease, with a step from the set {-2.0, -1.0, -0.5, -0.25, 0.25, 0.5, 1.0, 2.0}.

Fraction	Degrees	Orb *average orb	Weight
1/1	0	7.5 *	1
1/2	180	5.67 *	1
1/3	120	4.67 *	2
1/4	90	5.33 *	8
1/5	72	4.83 *	3
1/6	60	1.42 *	1.5
1/7	**51.42857**	**0.25**	**-1**
1/8	**45**	**1**	**-1**
1/9	**40**	**0.5**	**-8**
1/10	**36**	**0.5**	**-11**
1/12	**30**	**1**	**-1.25**
2/5	144	1	0.5
2/7	102.857	0.25	11
2/9	80	0.25	5
3/7	**154.2857**	**0.25**	**-4**
3/8	**135**	**0.75**	**-12**
4/9	**160**	**0.5**	**-0.75**
5/12	**150**	**0.75**	**-15.5**

Table 2. Orbs and weights of aspects, sorted by fraction: numerator then denominator

As we can see in Table 2, weights are positive for fractions 1/X, X=1...6, and fractions 2/Y, otherwise they are negative. If aspects are sorted by Degrees (Table 3), then signs of weights appear to be grouped too, but not as much as in Table 2, seven groups rather than four.

We tried other aspects, 3/10 and N/11, N=1..5, they did not improve the PO target, i.e. zero weights.

Degrees	Orb	Weight
0	7.5*	1
30	**1**	**-1.25**
36	**0.5**	**-11**
40	**0.5**	**-8**
45	**1**	**-1**
51.42857	**0.25**	**-1**
60	1.42*	1.5
72	4.83*	3
80	0.25	5
90	5.33*	8
102.857	0.25	11
120	4.67*	2
135	**0.75**	**-12**
144	1	0.5
150	**0.75**	**-15.5**
154.2857	**0.25**	**-4**
160	**0.5**	**-0.75**
180	5.67*	1

Table 3. Orbs and weights of aspects, sorted by degrees.

Here are the scaled Logistic Regression coefficients, averaged across 3000 iterations:

-100.0 Moon

47.93 Venus

45.82 Mars

This suggests that the most important feature is “Moon”, and this agrees with table 12 in [7].

Last but not least, our implementation[8] reports not only AVaSA and median VaSA, also the average TeSA for the 1500 cases where VaSA is above median, and the TeSA corresponding to the case where VaSA has the highest value among all the 3000 cases. For the main pair (SC, SMD) these values are as follows:

ATraSA=0.5770566

AVaSA=0.5973834

ATeSA=**0.5660876** = (0.5792243+0.5529509) / 2

medianVaSA = 0.5975100, mean TeSA when VaSA is above median = **0.5664914**

maxVaSA = 0.6167239, that TeSA = **0.5682513**

As we can see, ATeSA < avg. TeSA at top half VaSA < TeSA at the maximum VaSA.

If we consider all other pairs such that both Internal Testing sets contain at least 100 persons, we see that these inequalities hold in 90% of cases:

Pair	Size of Testing set	ATeSA	medTeSA	maxTeSA
SC, SMD	575 + 1364	56.60876	**56.64914**	**56.82513**
SC, Military men	541 + 1195	52.16316	**52.24642**	**52.37210**
SMD, Heredity vol. B	880 + 3150	55.12693	**55.14721**	**55.27002**
SMD, Heredity vol. E2	1332 + 760	53.97646	**54.00699**	*53.80492*
SMD, Mental patients	2018 + 177	58.10249	**58.53901**	**58.69981**

Table 4. Average TeSA at top half VaSA (medTeSA), and TeSA at the maximum VaSA (maxTeSA)

Results with Random Forest classifier

When we tried using the Random Forest function from the Scikit-learn library with all the default parameters, *classifier = RandomForestClassifier(),* ATraSA was higher than 98%, supposedly primarily because of **min_samples_leaf=1** by default[13]. We used **min_samples_leaf=100** instead.

In short, with Random Forest instead of Logistic Regression, ATraSA = 58.64% for the main pair (SC, SMD), but then AVaSA = 57.19%, and the Internal Testing set accuracies are only 58.05 and 52.18%. The External Testing set accuracies are 57.45 and 50.24%. Interestingly, the medTeSA-maxTeSA inequalities hold in 5/6 of cases, and maxTeSA is above 56.5% in all cases, see Table 5.

Testing set	Size	ATeSA	medTeSA	maxTeSA
Internal SC+SMD	575+1364	55.11803	**55.14748**	**56.53615**
External SC	235	57.44809	*56.88113*	**59.14894**
External SMD	51	50.24052	**51.37255**	**56.86275**

Table 5. ATeSA, medTeSA and maxTeSA with Random Forest used for training and evaluation.

Results with both groups from other AG professional groups

There are six other professional groups in AG with at least 1000 persons per group: actors, writers, journalists, musicians, painters, politicians. But as you can see in Table 6, there is no pair that would allow at least 1000 persons in the Training set, and at least 25 in each of the two Testing sets.

Group A, Group B	Training set size	Validation set size from group A	Validation set size from group B	Testing set size from group A	Testing set size from group B
Actors, Journalists	1734	669	149	224	2
Actors, Musicians	1238	954	705	187	6
Actors, Painters	1870	797	694	28	23
Actors, Politicians	2424	510	557	38	7
Actors, Writers	2440	508	436	32	9
Journalists, Musicians	740	596	474	52	486
Journalists, Painters	1334	348	443	3	542
Journalists,Politicians	1924	53	629	3	185
Journalists, Writers	1908	63	502	1	209
Musicians, Painters	2034	312	557	1	78
Musicians, Politicians	1402	628	1020	1	55
Musicians, Writers	1388	633	896	3	75
Painters, Politicians	2080	602	732	10	4
Painters, Writers	2066	599	623	20	9
Politicians, Writers	2862	339	223	6	11

Table 6. Training, Validation, Testing set sizes in pairs of other AG professional groups

So we decided to merge six groups into three: Writers and Journalists, Actors and Politicians, Painters and Musicians. Even if one fusion is totally unreasonable, all three pairs are fairly reasonable: Group1 versus Group2, Group1 versus non-Group1, Group2 versus non-Group2.

Furthermore, the same fusions are seen in AG themselves:

Volume A4 contains only Painters and Musicians (PM),

Volume A5 contains only Actors and Politicians (AP),

Volume A6 contains only Writers and Journalists (WJ).

Table 7 shows Testing set accuracies for the three pairs of three merged groups. All testing sets are small, but surprisingly maxTeSA is higher than 53% in all three cases.

Pair	Size of Testing set	ATeSA	medTeSA	maxTeSA
PM, AP	1 + 17	72.63725	79.23737	82.35294
AP, WJ	32 + 2	52.79687	53.16025	53.12500
WJ, PM	6 + 21	37.60000	38.44841	59.52381

Table 7. Testing set accuracies for the pairs of merged groups, as per volumes A4, A5, A6 in AG.

Investigation attempts

We intentionally do not report any p-values, but overall our results look like the null hypothesis is false. Therefore we designed and completed six groups of investigation experiments.

1 – Other random seeds. The Training/Validation splitting and then the outcome of our classifier depends on pseudo-random number generator, and the output of PRNG depends on the random seed. Therefore we checked performance on Testing sets with six other random seeds. Results are in Table 8 below. Note the Parameter Optimization process was applying the default seed, 321.

PRNG seed	SC internal	SMD internal	SC external	SMD external
321	57.92243	55.29509	57.57418	58.36013
1	57.94597	55.29734	57.57333	58.31242
12	57.94957	55.28695	57.59773	58.37059
123	57.94017	55.30005	57.57773	58.33791
1234	57.93490	55.31107	57.61376	58.35425
12345	57.95304	55.29599	57.61220	58.35425
123456	57.93687	55.28338	57.57730	58.33399

Table 8. Average Testing set accuracies with other PRNG seeds.

2 – Randomization of features. If the amount of duplicates and "**twins"** in the Training+Validation sets was quite big, that could become the main focus for what classifier would learn, and the main reason for high AVaSA, rather than the parameters being tuned (during the Parameter Optimization process) to the peculiarities of the Training+Validation sets. Duplicates are pairs of persons with the same moment of birth, and by "**twins**" we mean persons with very close moments of birth.

Our features are quite continuous: they are calculated in such a way that the closer the two moments of birth, the closer the values of features, on average. So if PO process and/or classifier learned mostly from duplicates, twins, and nearest neighbors, then the classification accuracies would remain high even if we applied a pseudo-random transform on each set of features, a transform that

would keep the features continuous. We tried six such transforms, details are in Table 9 below, and we observed the Testing set accuracies being not as high as without any pseudo-random transforms. Please see the main source file in our repository, namely MLAG_main.py, for more details.

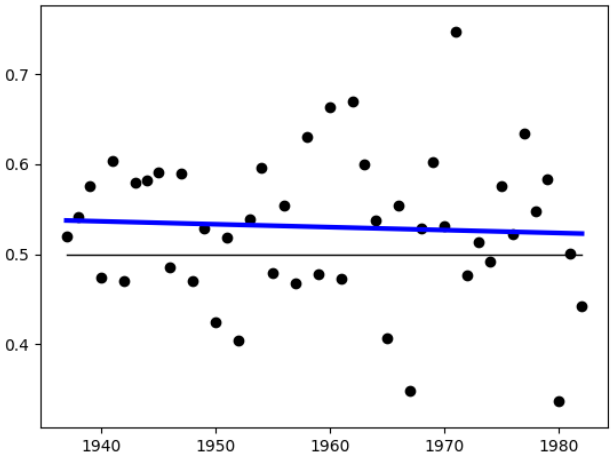
Randomization addendum	SC internal, 575 persons	SMD internal, 1364 persons	SC external, 235 persons	SMD external, 51 persons
379 + i*37	**46.87722**	52.39695	**48.44142**	57.43660
379 + i*53	**48.11641**	52.50758	**45.02567**	**42.67712**
379 + i*67	**47.54997**	**50.54626**	**48.14894**	**44.41438**
642 + (9-i)*41	**48.22017**	**49.27700**	**49.38922**	**46.73529**
642 + (9-i)*61	53.74806	**46.42351**	53.85220	**41.55229**
642 + (9-i)*79	**49.98986**	**44.49330**	**48.63319**	**39.57124**

Table 9. Testing set accuracies after pseudo-random transforms on sets of features.

The classifier consistently reports "*No, on average these persons are not like the Sports Champions from Training sets"* about SC's from both Internal and External Testing sets. Similarly "*No, they* are *not like Scientists and Medical Doctors from Training sets*" about SMD's from both Testing sets.

3 – Yearly averages. It is possible that there is a long-term trend in the data, something like "the closer to the end of the 20th century, the more likely it is that the person is a Sports Champion". Our three combined features, even though they were designed to avoid such trends, still could capture such trends to some extent. But in this case "Mars" would have a lot more chances than "Moon" to become the most important feature (Jupiter, Saturn even more). For further investigation we decided to check the percentages of moments in 1937-1982, with 1 hour step, such that a person born on that moment would be classified as a Sports Champion. The range is 1937-1982 because SMD's in the External Testing set were born in 1937-1968, and External SC's in 1958-1982.

We discovered that percentages of "SC" moments vary between 33.68% in 1980 and 74.73% in 1971. When we fit a linear regression, slope = -0.00032, see Figure 1, and the predicted values go from 53.74 to 52.28%.



In the preceding 46 year long period, 1891-1936 inclusive, the percentages of “SC” moments vary between 28.28% in 1905 and 70.51% in 1892. When we fit a linear regression, slope = 0.00015, and the predicted values grow from 52.69 to 53.35%.

When we re-run the same with step = **10 minutes** rather than **1 hour**, we see that the average percentages of “SC” moments are approximately the same:

53.0013 vs 53.0199 for the period 1891-1936 inclusive, and

53.0067 vs 53.0093 for the period 1937-1982 inclusive.

4 – Monthly frequencies. It is possible that Sports Champions were born more often during one season, and Scientists and Medical Doctors during another season. At least one of our three features could capture that, to some extent: “Venus” correlates with season, because the ecliptic longitude of Venus is equal to *the ecliptic longitude of Sun plus or minus 49 degrees or so*. Table 10 shows the frequencies and deviations for **Sun in the twelve 30-degree sectors starting from the vernal equinox**. We decided to use twelve astronomical “seasons” rather than the calendar months, as it seems more likely that moments of birth would correlate more with astronomical “seasons”. We merged Internal and External Testing sets for this experiment.

	Sports Champions, 810 persons			Scientists and Medical Doctors, 1415		
Sector number	Number of persons	Deviation, % of the total population	Deviation, standard deviations	Number of persons	Deviation, % of the total population	Deviation, standard deviations
0	63	-0.556	-0.523	123	0.359	0.402
1	52	-1.914	**-1.803**	125	0.501	0.560
2	75	0.926	0.872	107	-0.771	-0.863
3	64	-0.432	-0.407	87	-2.185	**-2.444**
4	76	1.049	0.989	117	-0.065	-0.072
5	73	0.679	0.640	122	0.289	0.323
6	67	-0.062	-0.058	135	1.207	1.350
7	72	0.556	0.523	111	-0.489	-0.547
8	58	-1.173	-1.105	111	-0.489	-0.547
9	81	1.667	**1.570**	124	0.430	0.481
10	56	-1.420	-1.338	117	-0.065	-0.072
11	73	0.679	0.640	136	1.278	1.429

Table 10. Sun in twelve 30-degree sectors: ecliptic longitudes of Sun, quantized to 12 values.

When we look at twenty four 15-degree sectors, we see that all deviations are in the range (-2.446, +2.009) standard deviations.

5 – Monthly frequencies, taking birthplace into account. This is an improved version of the previous investigation attempt. The majority of persons in AG were born in Western Europe, and the majority of those born outside of Western Europe were born in USA. It is possible that European-born Sports Champions were born more often during one season, while American-born more often during another season, but these anomalies could become invisible in the merged group with all SC's. We split all SC's from Testing sets according to the longitude of place of birth: those born in the Western Hemisphere are put into the "Western" subset, if the longitude of place of birth is to the West from 30°00′00″W.

Among SMD's from Testing sets too few, only 30, would be put to the "Western" subset, therefore we merge all SMD's from all AG and Testing sets into one set before splitting into "Western" and "Eastern" subsets. Note that for some persons the place of birth is unspecified.

Table 11 shows deviations for the case when SC's and SMD's are split into those born in the Western Hemisphere (with longitude of birthplace between 30°W and 180°W), and all other. We see that deviations are in the approximately same range (-2.4, +2.1) standard deviations.

	Sports Champions				Scientists and Medical Doctors			
	"Eastern" subset		"Western" subset		"Eastern" subset		"Western" subset	
Sector	Persons	Deviation	Persons	Deviation	Persons	Deviation	Persons	Deviation
0	39	-0.630	24	-0.165	308	0.132	8	-0.703
1	35	-1.317	16	-1.924	311	0.280	11	0.088
2	52	1.603	23	-0.385	313	0.379	17	1.670
3	39	-0.630	25	0.055	320	0.724	11	0.088
4	42	-0.115	34	**2.034**	304	-0.066	10	-0.176
5	45	0.401	28	0.715	314	0.428	13	0.615
6	42	-0.115	25	0.055	305	-0.016	13	0.615
7	43	0.057	29	0.934	272	-1.646	16	1.406
8	37	-0.973	21	-0.824	257	**-2.387**	5	-1.494
9	55	**2.119**	26	0.275	313	0.379	8	-0.703
10	37	-0.973	19	-1.264	310	0.230	4	-1.757
11	46	0.573	27	0.495	337	1.564	12	0.351

Table 11. Deviations, measured in standard deviations, for the "Eastern" and "Western" subsets.

6 – Hourly frequencies. It is possible, even if it seems unlikely, that for some demographic reason either Sports Champions or SMD's were born more often during a specific time of the day. It is unclear how our three combined features could capture that, but lets assume they could. In this case we would expect to see a significant excess or shortage not only in the hourly frequencies of moments of birth of SC's and SMD's (Table 12), but also an anomaly during the same hours in the data with percentages of "SC" moments (Table 13).

	Sports Champions, 810 persons			Scientists and Medical Doctors, 1415		
GMT hour	Number of persons	Deviation, % of the total population	Deviation, standard deviations	Number of persons	Deviation, % of the total population	Deviation, standard deviations
0	23	-1.327	-1.137	35	-1.693	-1.962
1	23	-1.327	-1.137	85	1.840	**2.133**
2	36	0.278	0.238	81	1.558	**1.805**
3	29	-0.586	-0.502	61	0.144	**0.167**
4	28	-0.710	-0.608	70	0.780	**0.904**
5	27	-0.833	-0.714	76	1.204	**1.396**
6	35	0.154	0.132	60	0.074	**0.085**
7	48	1.759	1.507	64	0.356	**0.413**
8	34	0.031	0.026	63	0.286	**0.331**
9	42	1.019	0.872	66	0.498	**0.577**
10	29	-0.586	-0.502	74	1.063	**1.232**
11	46	1.512	1.295	55	-0.280	-0.324
12	27	-0.833	-0.714	55	-0.280	-0.324
13	35	0.154	0.132	36	-1.622	-1.880
14	31	-0.340	-0.291	52	-0.492	-0.570
15	31	-0.340	-0.291	49	-0.704	-0.815
16	41	0.895	0.767	55	-0.280	-0.324
17	61	3.364	**2.881**	51	-0.562	-0.652
18	37	0.401	0.344	58	-0.068	-0.078
19	42	1.019	0.872	65	0.427	+0.495
20	31	-0.340	-0.291	48	-0.774	-0.897
21	22	-1.451	-1.242	55	-0.280	-0.324
22	17	-2.068	-1.771	49	-0.704	-0.815
23	35	0.154	0.132	52	-0.492	-0.570

Table 12. Frequencies of times of birth (GMT) quantized to 24 values.

	Years 1937..1982		Years 1937..1957		Years 1958..1982	
GMT hour	Percentage	Deviation, std	Percentage	Deviation, std	Percentage	Deviation, std
0	53.23618	1.375	52.56694	0.909	53.79834	0.941
1	52.94347	-0.386	52.27702	-0.240	53.50329	-0.277
2	53.37579	**2.214**	52.98343	**2.560**	53.70538	0.557
3	53.06409	0.340	52.38909	0.204	53.63108	0.251
4	53.27268	1.594	52.29944	-0.151	54.09019	**2.146**
5	52.82712	**-1.085**	51.90435	-1.718	53.60224	0.132
6	52.95872	**-0.294**	52.16490	-0.685	53.62552	0.228
7	52.77923	**-1.373**	52.24073	-0.384	53.23158	-1.399
8	52.79497	**-1.279**	52.35130	0.054	53.16764	-1.663
9	52.91048	**-0.584**	52.38090	0.171	53.35532	-0.888
10	52.76108	**-1.483**	52.30684	-0.122	53.14265	-1.766
11	52.86730	**-0.844**	52.11749	-0.873	53.49714	-0.302
12	52.99496	**-0.076**	52.46591	0.508	53.43935	-0.541
13	52.87384	**-0.805**	52.17281	-0.653	53.46270	-0.445
14	53.29399	1.722	52.39700	0.235	54.04745	1.970
15	52.91348	-0.566	52.16293	-0.693	53.54395	-0.109
16	53.05166	0.265	52.50276	0.655	53.51273	-0.238
17	53.01362	0.036	52.38428	0.185	53.54227	-0.116
18	53.10721	0.599	52.39053	0.210	53.70922	0.573
19	52.85644	-0.909	52.15481	-0.725	53.44581	-0.514
20	53.00575	-0.011	51.94947	-1.539	53.89301	1.332
21	53.05721	0.298	52.07977	-1.022	53.87826	1.271
22	53.17558	1.010	52.95637	**2.453**	53.35972	-0.870
23	53.04786	0.242	52.50445	0.661	53.50433	-0.273

Table 13. Percentages of “SC” moments and their deviations from the averages, measured in standard deviations. We decided to look at percentages of “SC” moments not only for the period 1937-1982, because all of the SC’s from the External Testing set were born in 1958-1982, while the vast majority of SMD’s in the External Testing set, 48/51 of them, were born in 1937-1957.

In Table 12 we see a small anomaly: more SMD’s were born between 0:30am and 10:30am GMT.

In Table 13 we see a somewhat similar, but an even smaller anomaly in the main column, namely 1937..1982 deviations: the percentages of “SC” moments are below average, deviations are in the range (-1.483, -0.076) standard deviations. But the time range overlaps only partially with the small anomaly in table 12: between 4:30am and 13:30am GMT.

The anomaly that we see in Table 12 looks a bit more unusual if we look at “Eastern” and “Western” subsets of SMD’s, as we did in the Investigation-5 experiments: in the “Eastern” subset only the deviations corresponding to hours 1am..11am are positive. But again, all 24 deviations are in the narrow range (-1.809, 1.902) standard deviations.

Conclusion

The results we presented, together with our investigation attempts, look like either a statistical fluke, or something that needs an explanation.

In case they are regarded a statistical fluke, we suggest that until the ultimate nature of reality, and in particular, the ultimate nature of time, are unknown, the community should keep an eye on such flukes rather than discard them.

In case a further investigation is carried out, we suggest that target groups should contain persons with distinctive psychological or physiological features, rather than outstanding professionals.

References

1. Archives Gauquein. http://cura.free.fr/gauq/17archg.html

2. Theories of astrology. Discourse by Dean, Loptson, Kelly and seven others in *Correlation* 1996, 15(1), 17-52. http://www.astrology-and-science.com/A-theo2.htm

3. Ken Irving. Misunderstandings, Misrepresentations, Frequently Asked Questions & Frequently Voiced Objections About the Gauquelin Planetary Effects. http://planetos.info/mmf.html

4. Jenkins, Fischbach et al. Evidence for Correlations Between Nuclear Decay Rates and Earth-Sun Distance. http://arxiv.org/abs/0808.3283

5. David P. Rothall, Reginald T. Cahill. Dynamical 3-Space: Gravitational Wave Detection and the Shnoll Effect. https://arxiv.org/pdf/1307.7437.pdf

6. Robert D. Doolaard. Waves of Wars. http://cyclesresearchinstitute.org/pdf/cycles-history/CRI200602-Doolaard-WavesofWars.pdf

7. James Gunasekera. New Patterns in Gauquelin Data. https://vixra.org/pdf/1106.0036v1.pdf

8. All code written and used for this report. https://gitlab.com/MLAG-hub/mlag/-/archive/master/mlag-master.zip

9. Steinbrecher Astrological Data Collection. https://www.astrocye.com/category.htm

10. SADC (Steinbrecher Astrological Data Collection) downloadable from AstroSystem.ru. http://astrosystem.ru/AstroSystem/Main/Research/Base/SADC/sadc.rar

11. SADC downloadable from the ZET web site. http://astrozet.net/zip/DBase/SADC.exe

12. Scikit-learn machine learning library. https://scikit-learn.org

13. Scikit-learn, documentation page on Random Forest Classifier. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html