A big data system for the classification of physical activity level in the Italian regions

Luigi Arminio[1], Jacopo Clocchiatti[2]

[1]luigi.arminio@studenti.unitn.it

[2]jacopo.clocchiatti@studenti.unitn.it

Abstract— Data about sports activity have often been collected through survey methods, which, however, have some critical issues (e.g., they are subject to social desirability bias, and they do not allow daily monitoring of physical activity in a given territory). In this sense, data registered in mobile applications, as well as OpenStreetMap data and Google searches data, if integrated with data collected in more traditional ways, could represent a way to have more detailed information about people's physical activity level. For this reason, in this work, we projected a big data system to classify the physical activity level in the different Italian regions based on multiple data sources. The output of the system is a daily updating DataBase containing several indicators, and a composite statistical index, to monitor the physical activity level in the Italian regions.

***Keywords*— Sport, Physical Activity, Jupyter, Python, Pandas, Apache Kafka, PySpark, MongoDB, Big Data**

1. Introduction

Traditional ways to collect data about physical activity include questionnaires and surveys. They are undoubtedly helpful in increasing knowledge about people's physical activity level, but these data can often be biased. For example, cognitive mechanisms like social desirability bias[1], perception distortions, and other factors can produce data that do not reflect reality. Furthermore, surveys do not allow daily monitoring: conducting surveys on a daily basis would be resource-intensive and time-consuming. In this sense, digital data that people spontaneously release through technological devices are a good way to enrich collectible information about the physical activity of people, constantly over time. For this reason, we decided to design a big data system that relies on multiple data sources and allows us to classify people's physical activity levels. Data sources that were thought to be functional to our goal are

1. Mobile apps data that count the number of steps or the repetitions of an exercise done by people during the day (Fitbit, Google Fit…)
2. Data from OpenStreetMap, which give us information about the number of sports facilities present in the territory
3. ISTAT data[2] about the frequency with whom people do physical activity in the different regions, collected through more traditional methods (survey, interviews)[3]
4. Sports-related Google searches data, that give information about the diversity of the frequency of sports-related searches in the Italian regions
5. Data about the number of accesses to the gyms of the different Italian regions, potentially obtainable from the gym turnstiles

Since mobile app data (1) and gym accesses data (5) were inaccessible to us for privacy-related concerns, we decided to simulate them based on a physical activity score determined through ISTAT data (3) exploration and transformation.
Mobile apps (1) can generate data about the daily steps of a person (which regard cardiovascular activity), as well as the number of squats (that refers to lower body exercise) and pushups done (which regards upper body exercise), but also data about the city in which a person is located.
OpenStreetMap data (2) lets us know how many sports facilities are in a given territory. With these data, we can register the number of gyms per 100000 inhabitants for each Italian region.
Google data (4) allows us to collect information about the frequency with which people search the word “sport” on Google. This is thought to give us information about people's interest in sports and can be considered an indirect measure of sportiness.

1.1 *Big data: The 5 Vs in our system*

In this paragraph, the data mentioned above are described according to the 5 Vs definition of Big Data:

- Velocity: Velocity: Data from mobile apps (1) are supposed to be daily reports regularly sent to the system (batch processing); Data from Openstreetmap are also processed in batch (monthly update of the number of sports facilities per region); Google search data are

scraped daily (batch); ISTAT data, used for the simulation, are static

- Volume: mobile apps generate millions of data points daily, which is a huge quantity. In the case of Google data, they are considerably small (20 data points, each referred to a region, released per day). Openstreetmap data and ISTAT data are small.
- Variety: Data from OpenStreetMap are in a structured form, while mobile apps data and gym accesses data are in JSON format (they are semi-structured)
- Veracity: Data from mobile apps (1) are simulated and, in this sense, not reliable. Hypothesizing to have real app data, they would have the pro of reducing the risk of social desirability bias because of a higher spontaneousness with whom data are released (steps are sensor-based, and squats and pushups are detectable through the phone camera). However, mobile apps, concerning ISTAT data (3) (which go through a careful sampling for the representativeness of the data) are more at risk of selection bias, to the extent that the population of sports apps users is different from the general population and the different distribution of technological standards in the different regions could affect the results. As for Openstreetmap data (2), they are accurate in detecting sports facilities in the territory, but the presence of several facilities in a region can be influenced by the industrialization value of that region and is not always a sign of higher sportiness. Google data (4) are directly scraped from Google and, in this sense, reliable in determining the search interests of the people; they somehow measure people's interests in sports, but interests do not always coincide with practice. Integrating data from several different sources may help to mitigate the effect of biases of specific sources.
- Value: All these data allow to build an automated system that monitors physical activity level in Italy at regional level. This allows to orient public policies in terms of interventions in the regions where physical activity is scarce.

1.2 *Expected output*

The aim of this project was to design and implement a data architecture that ingests data from a broad range of sources to classify the level of sportiness in a given territory. We decided to build this system for the Italian context to measure the level of sportiness in the different Italian regions, with constant updates over time.

The data source used to simulate data from mobile apps and gym ticket barriers was an ISTAT dataset containing data about people's sports habits, collected through surveys and qualitative interviews. For each region, the population is proportionally distributed into 4 categories that are hierarchically ordered based on their level of physical activity:

1) People who never practice sports activities (c1),
2) People which practice some type of physical activity (c2),
3) People that occasionally practice physical activity (c3),
4) People that continuously do physical activity (c4).

```
#Visualizing our data
istatdata.tail()
```

	ITTER107	Territorio	TIPO_DATO_AVQ	Tipo dato
571	ITD2	Provincia Autonoma Trento	3_NEV_SPORT	non praticano sport, né attività fisica
572	ITE1	Toscana	3_NEV_SPORT	non praticano sport, né attività fisica
573	ITE1	Toscana	3_NEV_SPORT	non praticano sport, né attività fisica
574	ITE	Centro	3_SPORT_OCCAS	in modo saltuario
575	ITE	Centro	3_SPORT_OCCAS	in modo saltuario

Fig. 1 ISTAT Data visualization on Pandas with Jupyter Notebook

For each region, the score is defined by weighting each of the 4 categories from 0 to 3, based on the physical activity level of each category. The first category (no sport) has a weight equal to 0, the second one has 1, and so on, up to the fourth one that has a weight equal to 3. These weights have been used to compute a weighted sum of the proportions of the different categories. Considering *c1…c4* as the proportions of people in the 4 categories and *w1…w4* as the respective weights, the sum is:

$$\textbf{\textit{wsum(pop)}} = c1*w1 + c2*w2 + c3*w3 + c4*w4$$

After that, the ISTAT-based sportiness score is generated by performing, on *wsum(pop),* a min-max normalization, based on the maximum (3*100 = 300 when 100% of people is in the fourth category) and the

minimum (0*100 = 0, when nobody is practicing sports) values possible in the previous sum:

ISTAT-based sportiness score = (*wsump(pop) - 0*) / (*300 - 0*)

For example, in 2021 Molise had 46.2% of the population from the first category (never practicing), 30.2% from the second one, 8.2% from the third one and 15.4% from the fourth one.
Hence, the computation of the weighted sum for Molise is:

*(46.2*0) + (30.2*1) + (8.2*2) + (15.4*3) = 92.8*

After that, it has been normalized:

(92.8 - 0) / (300 - 0) = 92.8/300 = 0.309

So, the ISTAT-based sportiness score of Molise in 2021 was 0.309 (0.31), which indicates low physical activity (2021 mean was 0.42).
This score has been computed for 2020 and 2021, and then the results have been averaged, obtaining the *Mean ISTAT-based sportiness* (MIbs) *score*, and the standard deviation has also been saved.

This MIbs score was used to weight the estimated proportion of people accessing gyms and to determine the mean activity detected by mobile apps (squats, steps, push-ups) through stochastic simulation. The simulation was performed based on the number of daily repetitions of squats, steps, and push-ups that are considered regular by previous research. Regular values were associated with 0.5 in the activity score (the intermediate value between high-level activity and no activity in the scale). Hence, considering R as the regular number of repetitions for any of the three activities, activity scores equal to 1 and 0 were conventionally associated to 2*R (2*0.5) and 0*R (0*0.5). So, each region's mean squats, steps, and push-ups were obtained by multiplying the MIbs score by (2*R)

2. System Model

2.1 *System architecture*

TABLE I
MOBILE APPS DATA - STRUCTURE

Variable	Type	Content
timestamp	int	Time at which data is reported
city_code	int	ISTAT code of the city in which the activity has been performed
type	str	Type of activity performed (steps, squats, push-ups)
activities	list	List composed of a tuple per device, each of them containing the device ID (str) and the n° of steps/squats/push-ups performed on the daily

To build our big data system, a pipeline composed of several stages was designed. It begins with the simulation of mobile apps data and gym accesses data; after that, there is the ingestion of data from multiple sources, followed by the data processing stage.

Data simulation can be considered the stage 0 of our pipeline. Since we did not have real mobile apps data or gyms entrances data, these data were simulated based on the previously defined mean ISTAT sportiness score (based on 2020 and 2021 data). By taking inspiration from the way in which data from sports apps like Fitbit and Google Fit are structured, we simulated the daily reports of the number of steps (to measure cardiovascular activity), squats (to measure the intensity of lower body workout in a person), and push-ups (to measure the intensity of upper body workout in a person). Based on scientific literature, we found a number of repetitions that represent a regular daily exercise for each of the three above-mentioned activities. We found, as regular daily repetitions for the three exercises, 5296 steps, 21 push-ups, and 39 squats. Since the MIbs score was in scale 0-1, where 0 means absence of activity and 1 means highly intensive activity, we considered 0.5 as a regular activity level. Hence, as written in the previous chapter, the above-mentioned repetition figures (5296 steps, 21 pushups, and 39 squats) are the regular values R attributed to the regions with MIbs score = 0.5. The doubled figures, considered the “maxima” (10592, 42, 78), were attributed to the regions with MIbs score = 1.0 (0.5*2). To simulate the mean number of daily repetitions per region, for each activity type, we multiplied the ISTAT score of each region by the maximum number of repetitions, and the result was rounded up. For example, in Lombardia (MIbs score = 0.4965), the mean number of daily squats is equal to 0.4965*78 ≈ 39. The same approach has been used for

each region, for all the activities. This allowed us to generate a JSON file per city containing the city code, the exercise type, and a list of tuples containing all the exercise data of the people from that city (the n° of steps or squats/push ups done i the day by each person was simulated through gaussian simulation, based on mean ISTAT score and its variance). The JSON files are sent everyday at the end of the day.
The MIBs score was also used to weight the proportion of population accessing the gyms from the different regions. In this case, a JSON with the accesses to the gyms of the different cities is daily generated. The simulation was essential to enable us to ingest data from several sources.
The pipeline stages that follow the simulation are the subsequent ones:

- *Data ingestion*: In this phase, data from several sources are collected and integrated: In addition to the simulated data from gyms and mobile apps, we collected, from Openstreetmap, data about the number of sports facilities in the different regions, and data about Sports-related Google searches data for the Italian regions.
- *Data Processing*: At this stage, ingested data go through data cleaning and transformation. Mobile apps data and gym accesses data are aggregated and averaged by region (in other words, the mean daily steps/squats/push ups and the proportion of people going to the gym for each region are computed).
  Then, for each item used to measure people's physical activity (mobile apps data, google searches, proportion of gyms in the territory, gym accesses), a copy of the item is scaled to a 0-1 range through min-max normalization. After that, all these scaled items are aggregated and averaged to generate a "Sportiness Index" on a scale 0-1 (from the least to the most sporty region), rounded up to the nearest cent.
- *Persistent Storage:* the "Sportiness Index" and the non-scaled (original) items mentioned before are selected to be stored and saved on a Database.

The final output is a database, in constant update, which records the mean daily steps, squats, and push-ups of people in the different regions, the number of gyms per 100000 inhabitants in each region, the frequency with whom people attend gyms in each region, the frequency of sports-related Google searches, and the Sportiness Index of each region.

.

A. Technologies

For the simulation of our data, the used programming language was Python. Python was also initially used with Jupyter Notebook for exploratory purposes, along with traditional Python packages. In particular, Pandas was used to explore ISTAT data and to define a MIbs score that was functional to the simulation of our data. Data about MIbs score and simulated values according to daily steps, squats, and push-ups of the people have been stored on a structured DB (with PostgreSQL). After that, NumPy was used to perform a data-driven stochastic simulation (based on the above-mentioned MIbs score) according to a Gaussian distribution through the *numpy.random.normal* function.

In the data collection phase, two APIs are used. Overpass was used to collect Openstreetmap data about sports facilities. To collect Google searches data, the used API was PyTrends, which is an unofficial API that scrapes data from Google Trends and has already been used in peer-reviewed research papers[4]. An issue with Pytrends data is that, since these data are obtained by web scraping, in case of future changes in the structure of Google Trends pages, data may not be generated correctly.

For the ingestion of data from multiple sources, since we were dealing with a huge volume of semi-structured data that are produced rapidly, we used MongoDB. PostgreSQL was also used to read Openstreetmap data. As for message queues, Kafka was preferred to other message queueing services because of its ability to manage high volumes of data.

In the data processing stage, PySpark was used to process data with parallel computing, according to the MapReduce paradigm. Even in this case, the choice of PySpark was related to its ability to manage huge volumes of data. PySpark allowed us to normalize our data and to compute the Sportiness Index

MongoDB was also chosen as a persistent storage tool; it was used to store aggregated and cleaned data about physical activity in the Italian regions on a daily basis.

3. Implementation

The code written to build this data architecture is accessible on a GitHub repository[5]. It is composed of a set of scripts that must be executed in separated shells:

-The *simulators* (*simulator_activity.py, simulator_facilities.py*), respectively used to simulate the intensity of physical activity and the accesses to the gyms

-The *consumer script* (*consumer.py*), which starts the phase of ingestion and message queuing (with MongoDB and Kafka)

-The *Spark script* (*sport_index_computation.py*), which starts the processing phase. This script allows to compute the Sportiness Index and to store the clean results of the processing phase (mean activities per region, sports facilities per region, people going to the facilities per region, Google searches per region, Sportiness Index per region) in a NoSQL Database, with MongoDB. The data, which are daily updating, are requirable through a front-end query to the MongoDB data.

4. Results

The results of our pipeline consist of a daily updating NoSQL Database containing, for each Italian region, information about the main items of sportiness (according to how discussed in this work) and the Sportiness index.

TABLE II
SPORTINESS OF THE ITALIAN REGIONS - FINAL DB

Variable	Type	Content
Sports Facilities	float	Number of sports facilities per 100000 inhabitants in a region
Research interest	int	Frequency of searches of the 'sport' word on Google
Daily steps	float	Mean daily steps done in a region
Daily squats	float	Mean daily squats done in a region
Daily push-ups	float	Mean push-ups done in a region
Sportiness index	float	Sportiness level of a region (0-1 scale)

By exploring the results, it turns out that the simulated mobile apps data, which actually reflects ISTAT data, shows that Physical activity is higher in the regions of Northern Italy (Trentino Alto Adige, Valle D'Aosta, Veneto and Lombardia are the regions with the highest values in terms of "mobile apps" data). On the contrary, southern regions, like Campania, Sicily, and Calabria, had the worst numbers.

As for Google searches, the situation appeared to be different. By querying daily searches, we found Lazio and Marche among the regions with the highest proportion of sports-related searches; also in this case, Sicily had a bad score.

In general, by looking at the overall index, data seems to suggest a higher sportiness in central-northen Italy.

5. Conclusions

The realization of this project starts from the assumption that traditional ways to investigate the sportiness of the population in a given territory can be integrated with cutting-edge methods based on the huge volumes of data constantly released by contemporary technological devices. However, we had to deal with the difficulty of finding "spontaneously released" data by common people (in this case, mobile apps data) because of privacy-related and copyright-related concerns. In this sense, it appears that developing new mobile apps to track people's daily activities can be crucial for research projects that aim to accurately track human behaviour by taking advantage of the current technological development.

As for the limitations of the project, it must be remembered that it only took into consideration the Italian territory on a regional basis (this is related to the fact that ISTAT data, which were crucial for the simulation, were region-based and not city-based)
Working with city-based data coming from the whole planet could need higher resources to process the data.
Furthermore, we implicitly assumed to be the holders of a mobile app that reports a code (based on ISTAT numbers) associated with the city where the activity occurs. However, mobile apps often return geographical data in terms of coordinates, which can need other APIs for reverse geocoding to obtain the name of the city where an activity occurs.

As for the possible enhancements for this project, the final MongoDB database may be integrated with an online Dashboard reporting both data about Sportiness in the Italian regions in a tabular format and a Choropleth map to visualize the "daily situation" in terms of sportiness in the Italian territory. However, since the data are reported in terms of region names, it would be necessary to use a GeoJSON file containing the coordinates of each region associated with the region names, or the use of tools like Tableau, which can automatically detect the coordinates of a region based on its name (when written correctly.)

References

[1] Brenner, P. S., & DeLamater, J. D. (2014). Social desirability bias in self-reports of physical activity: is an exercise identity the culprit?. *Social Indicators Research*, *117*(2), 489-504.

[2] http://dati.istat.it/Index.aspx?QueryId=24361

[3] https://www.istat.it/it/archivio/217037

[4] Zitting, K. M., Lammers-van der Holst, H. M., Yuan, R. K., Wang, W., Quan, S. F., & Duffy, J. F. (2021). Google Trends reveals increases in internet searches for insomnia during the 2019 coronavirus disease (COVID-19) global pandemic. *Journal of Clinical Sleep Medicine*, *17*(2), 177-184.

[5] https://github.com/jclock98/BDT-project