

Источники данных (сравнение на примере данных по Испании и Франции)

В сравнении участвуют четыре источника, правда, в разной степени детализации. Это данные университета Джона Хопкинса (№1): https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series; портал ЕС с доступными данными о COVID-19 (№2): <https://data.europa.eu/euodp/en/data/dataset/covid-19-coronavirus-data/resource/55e8f966-d5c8-438e-85bc-c7a5a26f4863>; данные Роспотребнадзора (docx-файл на каждый день), №3: https://www.rospotrebnadzor.ru/region/korono_virus/epid.php и данные с сайта **Внешняя Торговля России** (№4): <https://russian-trade.com/coronavirus/>.

Начнем с того, что Роспотребнадзор от 28 мая дает общее число инфицированных в Испании равным 236769, но указывает, что Испания не включает в себя 47080 случаев, подтвержденных серологическими методами (с помощью ПЦР), так что вместе с этими случаями общее количество инфицированных составляет 283769. Эти два значения вполне согласуются и с данными от источника №4, дающего число 283849 = 283769 · (1+0,000282) за 27 мая (Роспотребнадзор отстает на сутки), и с данными от Хопкинса, если учесть весьма частый для источника №1 эффект пропуска в заполнении данных. Поясним этот эффект и согласованность, используя следующий ряды данных:

Даты наблюдений	26 май	27 май	28 май	29 май
Общее число, Хопкинс	236259	236259	237906	238564
За сутки, Хопкинс	859	0	1647	658
За сутки, портал ЕС	859	510	1137	658

Легко видеть, что 0 за 27 мая от Хопкинса – очевидный пропуск одного дня, число инфицированных за который добавлен к 28 мая (1647 = 510+1137), так что 27 мая, если восстановить пропущенное значение, общее число будет как раз равно 236769. И таких пропусков в данных от Хопкинса не так уж мало. А теперь обратим внимание на следующую диаграмму (рисунок 1).

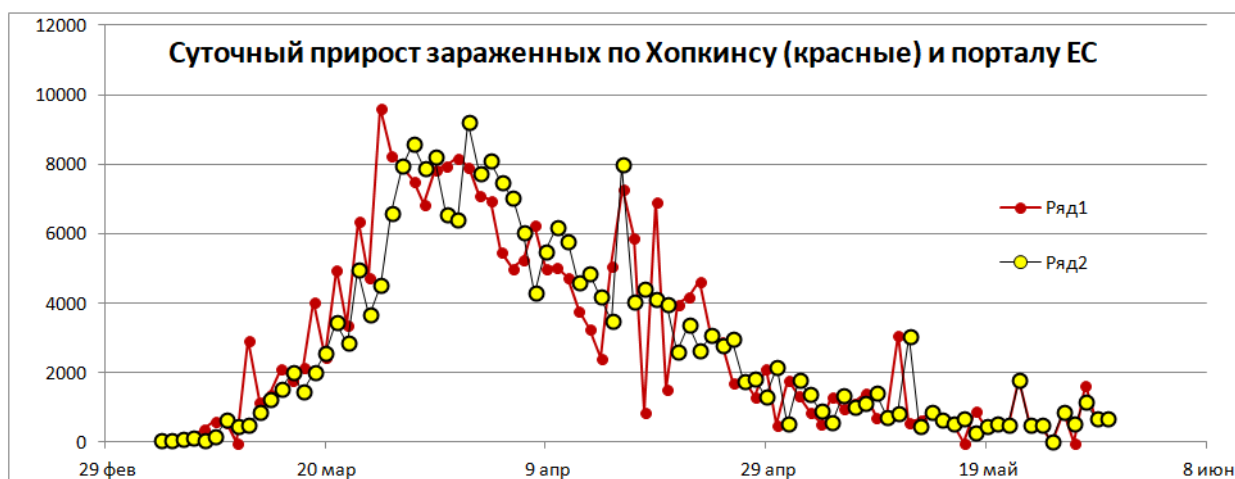


Рис.1. Это данные по динамике числа инфицированных за сутки от источников №1 и №2.

Примечание. Среди данных было два отрицательных значения: -10034 (24 апреля) и -372 (25 мая). Они убраны из рядов данных по причине их искусственности (пересчет зараженных).

На диаграмме везде, где видны красные точки, имеется расхождение в данных. Непомерно высокие выбросы вверх в марте в данных Хопкинса никак не согласуются ни с данными портала ЕС, ни с данными ВОЗ, ни даже с естественным ходом их собственной динамики: такая "пила" ничем не

может быть объяснена. Вызывают удивление и броски вверх-вниз в интервале 15-25 апреля.

За этим кроются не только естественные по своей природе расхождения в данных от разных источников из-за разной привязки к разным часовым поясам, но и в банальной небрежности сборщиков этих данных. Такого рода ошибки присутствуют и в данных ВОЗ, и на портале статистических данных ЕС, и далеко не редко в данных от университета Джонса Хопкинса. Вот ещё один пример – данные по накопленному числу инфицированных Covid-19 во Франции:

8 апр.	9 апр.	10 апр.	11 апр.	12 апр.	13 апр.	14 апр.	15 апр.
82048	86334	90676	93790	120633	124298	129257	132473
16 апр.	17 апр.	18 апр.	19 апр.	20 апр.	21 апр.	22 апр.	23 апр.
144944	146923	146906	151808	154188	156921	154715	157026

Казалось бы всё хорошо, но вычислим по этим данным суточный прирост инфицированных:

8 апр.	9 апр.	10 апр.	11 апр.	12 апр.	13 апр.	14 апр.	15 апр.
	4286	4342	3114	26843	3665	4959	3216
16 апр.	17 апр.	18 апр.	19 апр.	20 апр.	21 апр.	22 апр.	23 апр.
12471	1979	-17	4902	2380	2733	-2206	2311

И на протяжении всего 16 дней крайне странно выглядит **отрицательный прирост**, особенно на **-2206** человека, и огромный прирост почти в **27 тысяч** инфицированных на фоне 3-5 тысяч в остальные дни. Причем, насколько мне известно, Франция не пересчитывала число своих инфицированных, в отличие от Испании. Сравнение суточных данных числа инфицированных во Франции от университета Хопкинса и от портала ЕС представлено на рисунке 2.

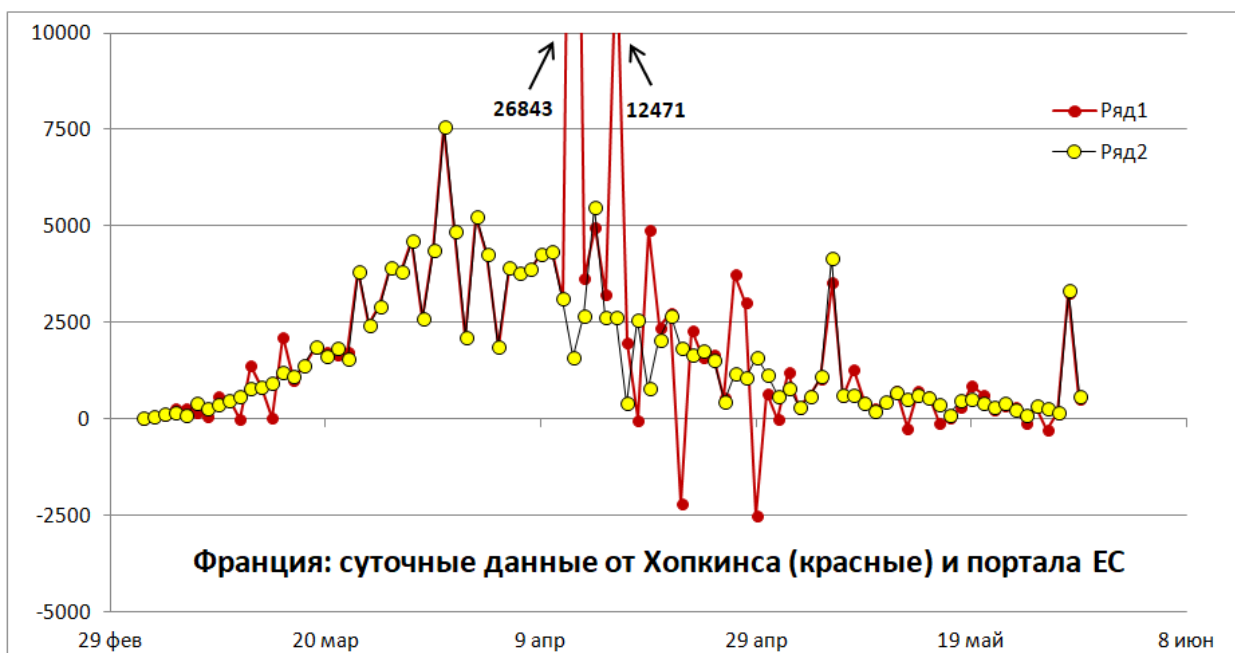


Рис.2. Данные по динамике числа инфицированных за сутки во Франции от ун-та Хопкинса (красный цвет) и портала ЕС. Везде, где видны красные точки, имеется расхождение в данных.

Опять-таки мы наблюдаем совершенно неправдоподобные выбросы вверх 12 и 16 апреля, уже отмеченные в выше приведённой таблице и скачки вниз, в область отрицательных чисел. Причем, в соответствующей базе с данными никаких пояснений об этом от составителей нет. Выводы из всего этого сопоставления данных пусть делает читатель.