Raspando a web sobre prática de atividade física

Marcos Duarte

em colaboração com

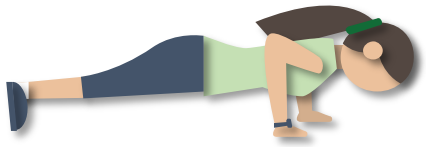
Leonardo Amaral Afonseca & **Renato Naville Watanabe**

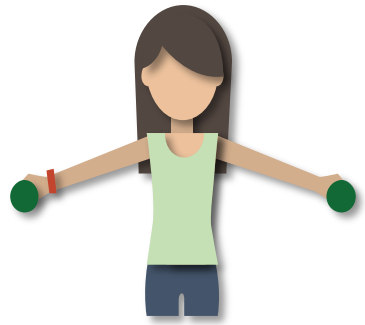
Laboratório de Biomecânica e Controle Motor

Universidade Federal do ABC

46º SIMPÓSIO INTERNACIONAL DE CIÊNCIAS DO ESPORTE – CELAFISCS (14/09/2023)

E se eu dissesse que é possível coletar registros diários da prática de atividade física por milhares de pessoas em todo o mundo?

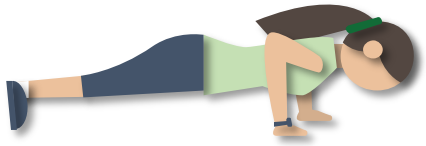


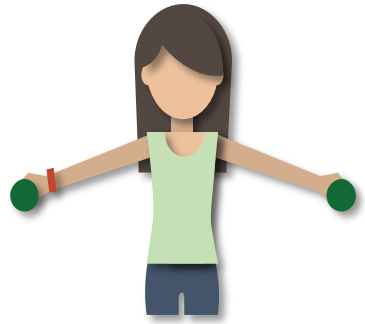


But Wait...

THERE'S MORE!

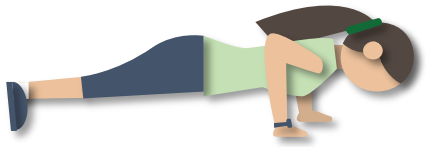
E se para esta coleta de dados você nem precisasse sair do laboratório?

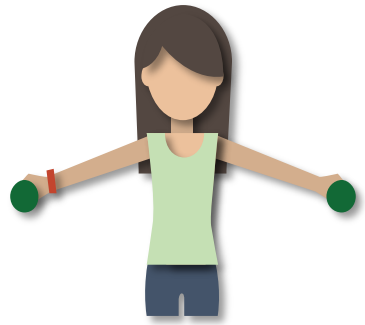





BUT WAIT,
there's more!

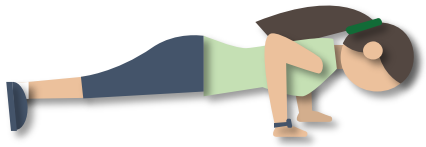
E se você conduzisse uma pesquisa com estes dados e nem precisasse submeter o projeto ao comitê de ética em pesquisa?

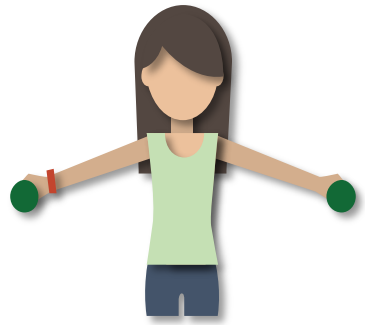





D.R.I.
ACT NOW
AND GET THESE OTHER GREAT ITEMS FREE
AWESOME PRODUCTS!
SKANKER
As Seen On TV
AGAINST ME
PERSONAL MANAGER
D.R.I.
COUCH SLOUCH
But Wait... THERE'S MORE!

E se toda a pesquisa fosse grátis?

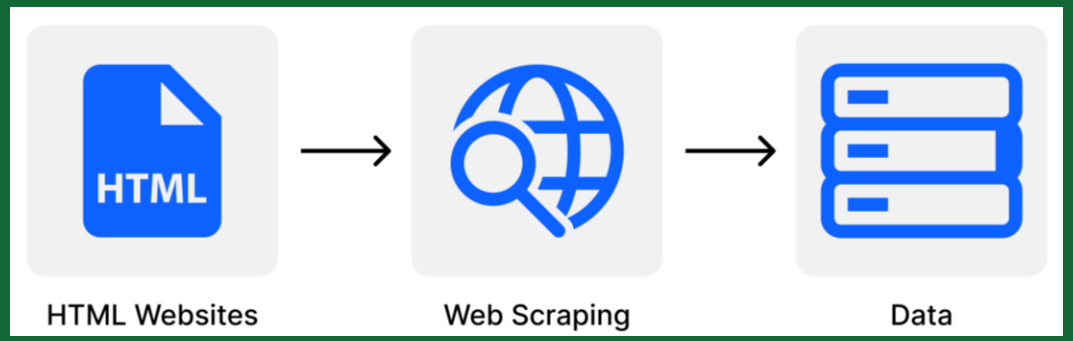


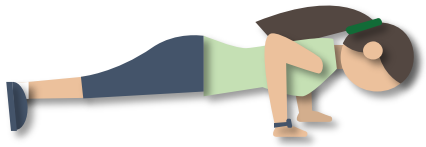


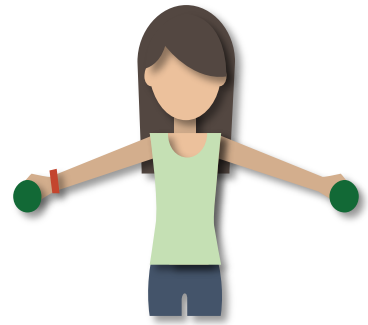
Apresentando:

Web Scrapping (raspagem web)

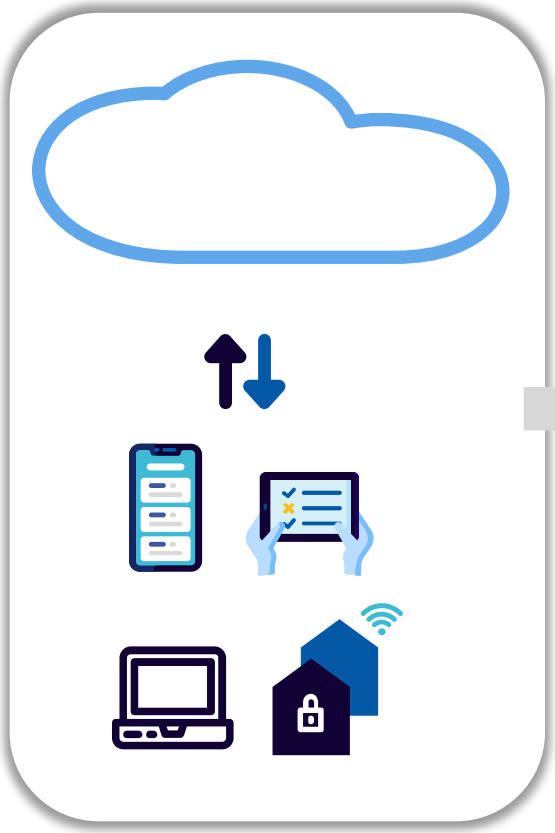
Uma forma de mineração que permite a extração de dados de sites da web convertendo-os em informação estruturada para posterior análise [Wikipedia].







Registros de atividade física na internet



Cada vez mais dispositivos se conectam à internet e potencializam a geração de dados

Acredita-se que tenhamos um aumento exponencial até 2025 no volume de dados

Empresas de diversos segmentos têm gerado valor a seus clientes através da entrega de produtos e serviços a partir de seus dados

STRAVA

1. Dados acessíveis e públicos
2. Grande volume de dados
3. Marca com grande crescente de usuários
4. Rede social esportiva
5. Hub de informações de muitas marcas de relógios esportivos

UFABC

01

Ferramenta para extração

Desenvolvimento de script para raspagem de dados da plataforma Strava

02

Dataset público

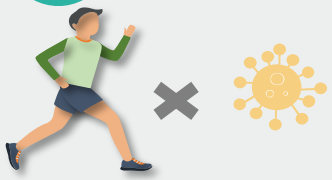
Disponibilização de forma pública de uma base de dados anonimizada

03

Análise exploratória das atividades

Análise exploratória das atividades dos atletas praticadas em 2019 e 2020

04

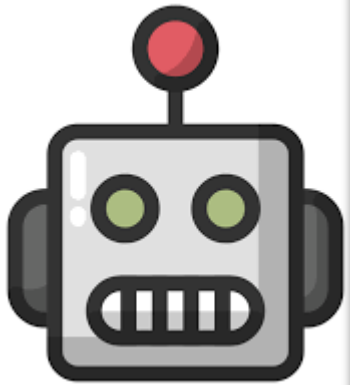


Covid-19

Análise quantitativa do impacto no volume de atividades físicas (corrida) realizada por maratonistas

UFABC

Ferramenta de Webscraping - Strava Robot



Reconectável

Para cada rotina no programa, a conexão é verificada, e no instante que se perde, o processo entra num ciclo de standby, e a cada minuto a internet é verificada, caso aconteça logoff, o login é refeito e o processo inicia do ponto que parou

Parametrizável

O usuário pode optar pelo número de atletas a serem extraídos, tal qual, período de aquisição.

Escalável

Através da biblioteca Papermill é possível paralelizar o processo, afim que seja feito o acionamento de notebooks em paralelo, otimizando assim a aquisição de múltiplos dados ao mesmo tempo

Autônomo

Caso seja identificado falhas no site, como não carregamento de informações, a página é atualizada até que não apresente mais falhas. Exemplo: manutenção do site, ou intermitência sistêmica

Banco de dados gerado a cada iteração

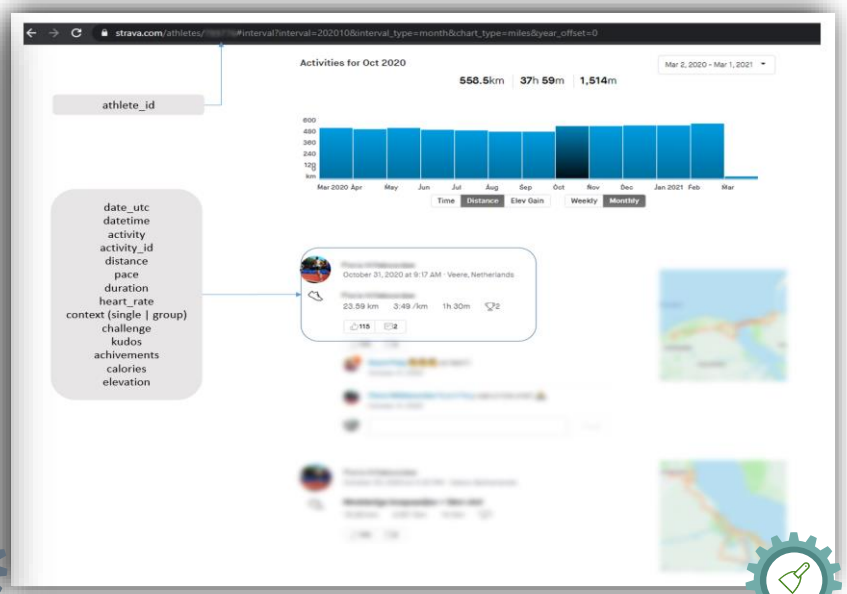
Ao fim da aquisição das informações de cada atleta, o banco de dados é atualizado, com o intuito de sempre manter a última versão, mitigando assim possíveis problemas referentes a falha na máquina do usuário, ou reset indevido durante o processo. Em cada notebook é possível acompanhar a evolução da extração dos dados através de **log em tela.**

Velocidade de aquisição

Lista de atletas: **1s / atleta (10h – tempo total)**
Lista de atividades: **5 min / atleta (3133h – tempo total)**
País de cada atleta: **3 min / atleta (1880h – tempo total)**

UFABC

Base de dados



- Padronização dos horários e datas
- Eliminação de dados inconsistentes
- Atividades de corrida / ciclismo com tempos superiores a recordes mundiais

Critérios de **inclusão**

- Atleta corredor, que tivesse corrido uma das 6 maratonas até 2019
- As extrações foram limitadas as atividades do período de Janeiro de 2019 a Dezembro de 2020

Critérios de **exclusão**

- Foram excluídos atletas que não compartilhavam suas atividades publicamente
- Atletas que não possuíam as informações de faixa etária

- Aquisição da lista de atleta; Seleção do sexo, faixa etária
- Iteração de cada página, e armazenamento dos dados; Aquisição da ID de identificação do usuário, do evento, com data e hora
- Iteração mês a mês, e armazenamento de todos os grupos de informações para cada atividade
- Aquisição dos detalhes das atividades (data e horário local/UTC, ID de identificação da atividade

- Os dados foram armazenados anonimamente sem a presença de IDs que possam identificar o atleta e/ou a atividade
- O armazenamento se deu de forma bruta, e reamostrados em função dos dias, semanas e meses
- As bases de dados foram disponibilizadas em arquivos tabulares (.parquet)

Base de dados

- A análise de dados foi feita sobre as **atividades de corrida**
- Os **intervalos de confiança** foram calculados por bootstrap com 100.000 amostragens aleatórias com repetição
- Execução de **análise exploratória**: atleta e atividades
- **Volume de treinamento semanal** e tempo dos atletas nas maratonas (segmentado em intervalos de 30 minutos)
- Análise do impacto da pandemia **COVID-19** no treinamento dos atletas (número de atletas que correram por semana, número de corridas por semana, distância de corrida semanal, duração de corrida semanal, e o ritmo médio semanal. Tais dados foram comparados entre 2019 e 2020.

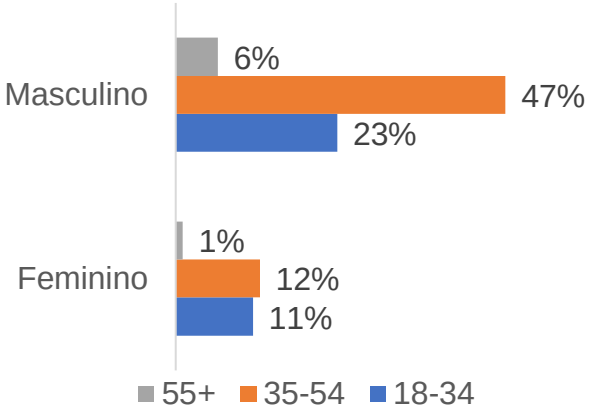
- Fechamento de escolas
- Fechamento de locais de trabalho
- Cancelamento de eventos públicos
- Restrições a reuniões públicas
- Fechamento do transporte público
- Ações para ficar em casa
- Restrições a viagens internacionais
- Campanhas de informação pública
- Restrições aos movimentos internos

Bloqueios governamentais e o Índice de Estringência

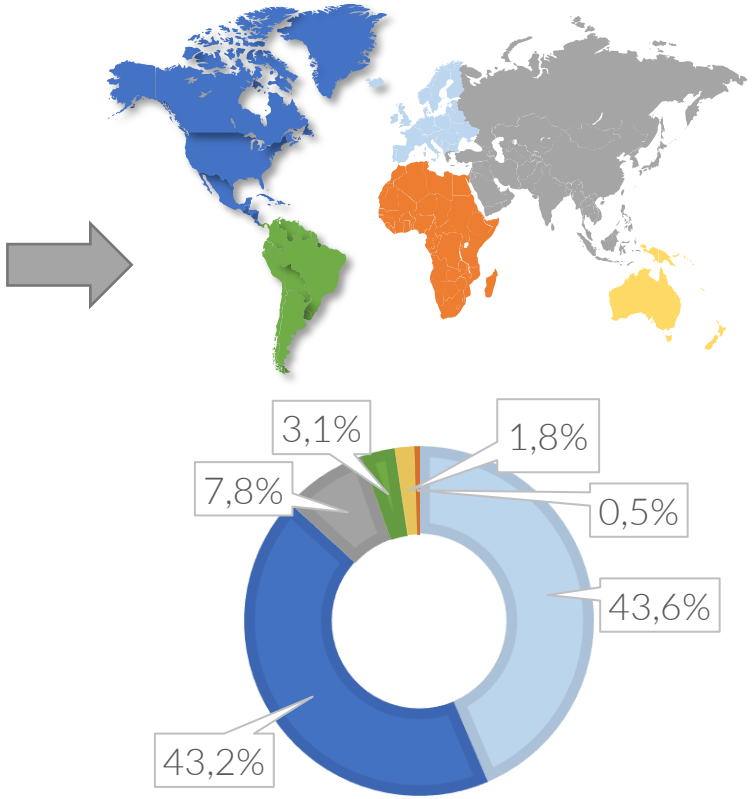
UFABC

Atletas

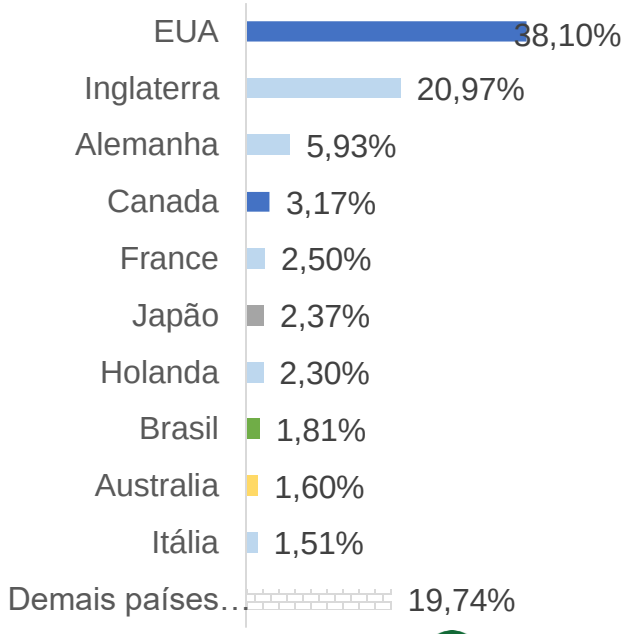
36k atletas



Divisão por continentes

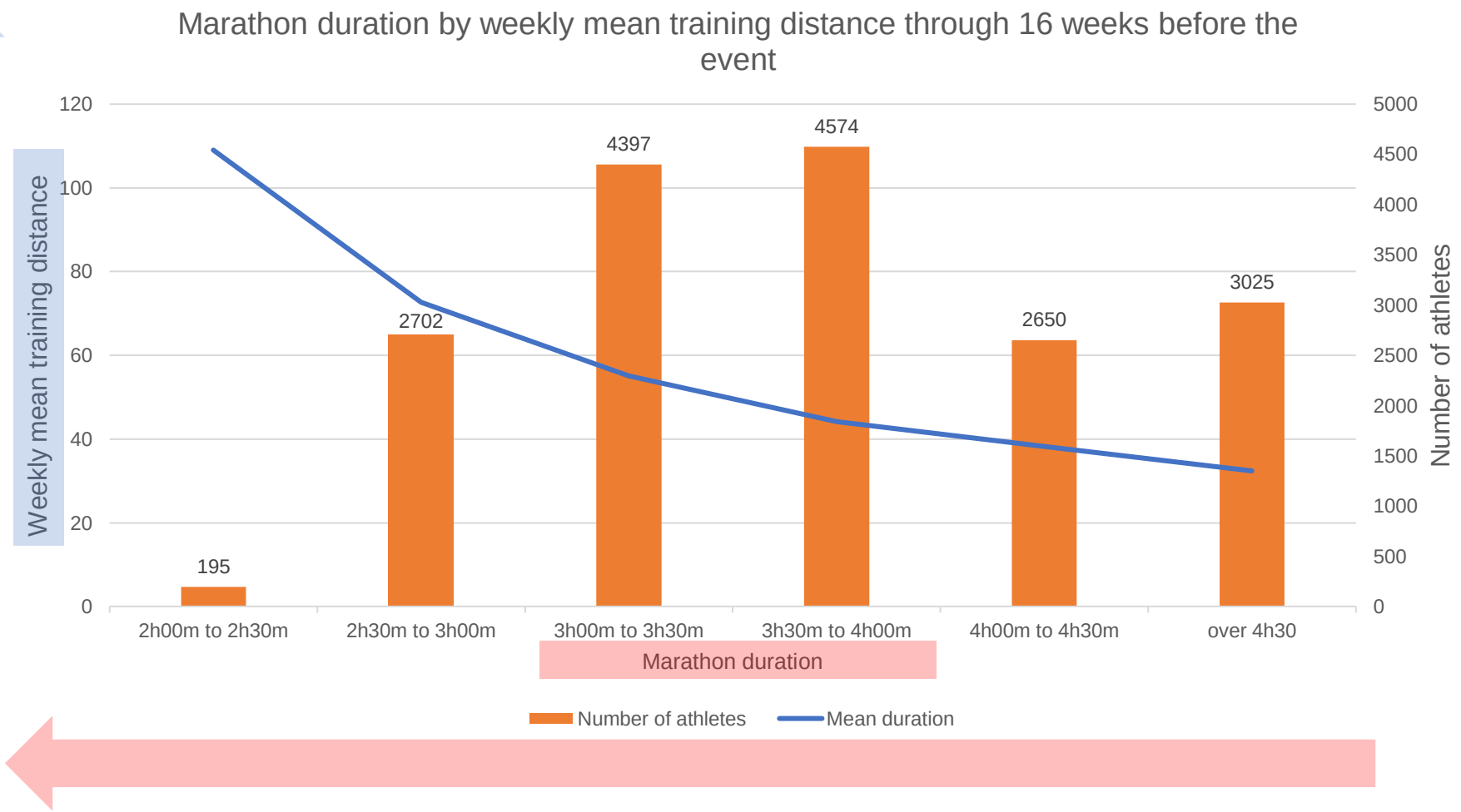


Divisão por países



UFABC

Volume de treinamento e tempo da maratona



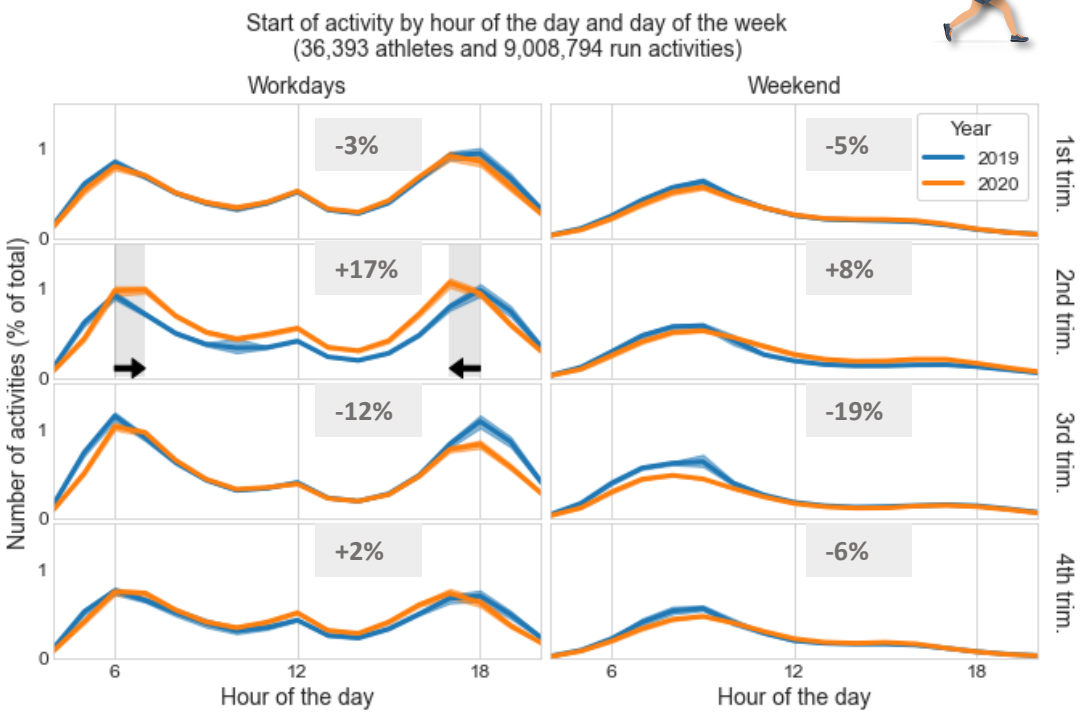
+17k
maratonas

17486
atletas

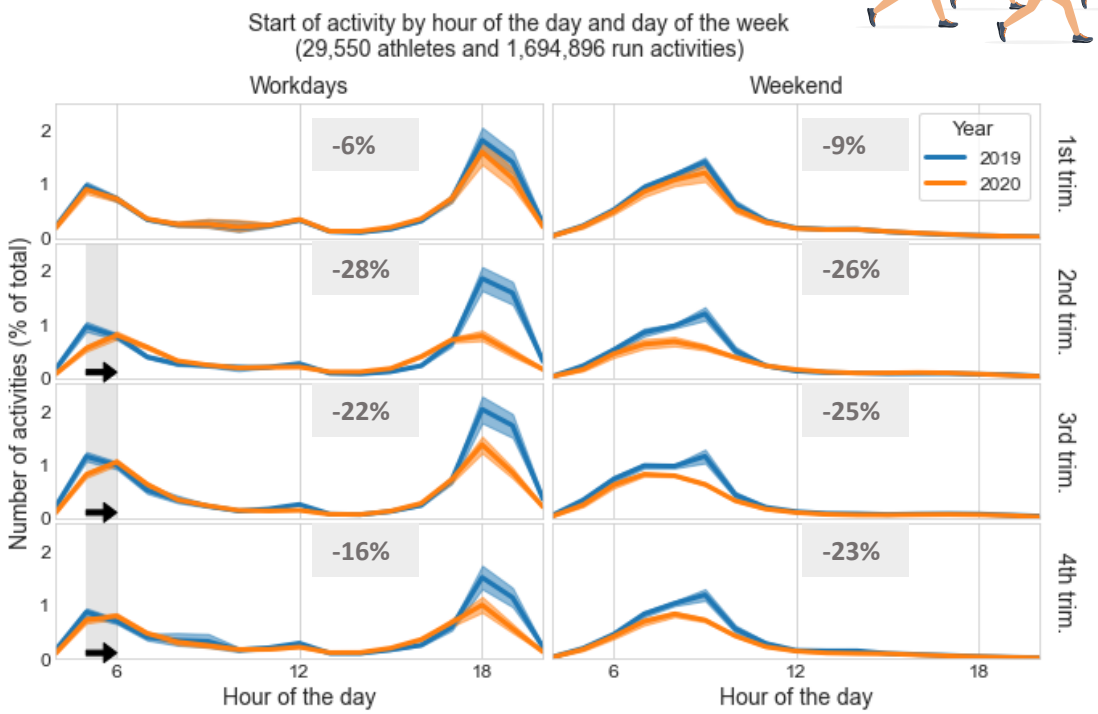
Tempo na maratona	F	M	Todos
2h00m to 2h30m	0,01%	1,11%	1,12%
2h30m to 3h00m	0,80%	14,62%	15,42%
3h00m to 3h30m	4,12%	20,94%	25,07%
3h30m to 4h00m	7,34%	18,72%	26,06%
4h00m to 4h30m	4,90%	10,21%	15,11%
over 4h30	6,83%	10,40%	17,23%
All	24,00%	76,00%	100,00%

UFABC

Efeitos associados do COVID-19 na prática de corrida de longa distância - Mundo



- Perfil similar da prática esportiva aos finais de semana (início aproximado as 9h)
- No terceiro trimestre, aos finais de semana, é apresentado a maior queda percentual no número de atividades
- O Segundo trimestre é caracterizado por uma mudança de comportamento dos atletas, com a mudança do início das atividades das 6h em 2019 para as 7h em 2020, e das 18h em 2019 para as 17h em 2020



- De uma forma geral as atividades tiveram queda acentuada, tendo maior impacto nos treinos ao final do dia (em dia de semana), e nos treinos diurnos (aos finais de semana)
- Maior queda observada em dias úteis no segundo trimestre
- As atividades em grupo a partir do segundo trimestre, passam por um mudança de comportamento dos atletas, com o início das atividades das 5h em 2019 para as 6h em 2020

±x% Diferença percentual da quantidade de atividades de 2020 em relação a 2019

UFABC

PeerJ Life & Environment

View 4 tweets

Related research

Share

‹ SPORTS MEDICINE AND REHABILITATION

A worldwide comparison of long-distance running training in 2019 and 2020: associated effects of the COVID-19 pandemic

Research article | Kinesiology | Public Health | Statistics | Data Science | COVID-19

Leonardo A. Afonseca, Renato N. Watanabe, Marcos Duarte

Published March 25, 2022

Highlighted in Coronaviruses and Viral Respiratory Infections

Author and article information

Biomedical Engineering, Universidade Federal do ABC, Sao Bernardo do Campo, SP, Brazil

DOI
10.7717/peerj.13192

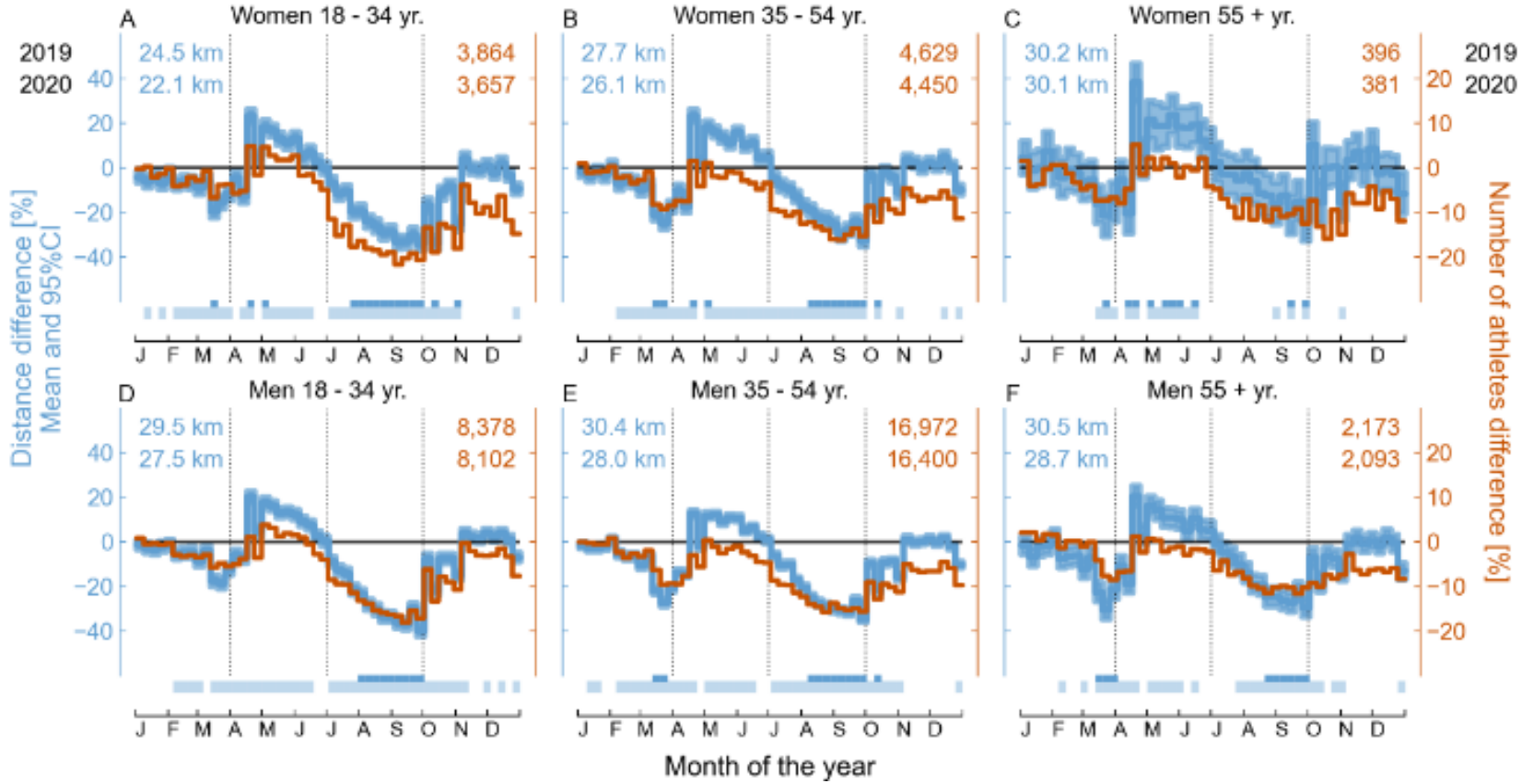
https://peerj.com/articles/13192/

UFABC

Efeitos associados do COVID-19 na prática de corrida de longa distância - Mundo

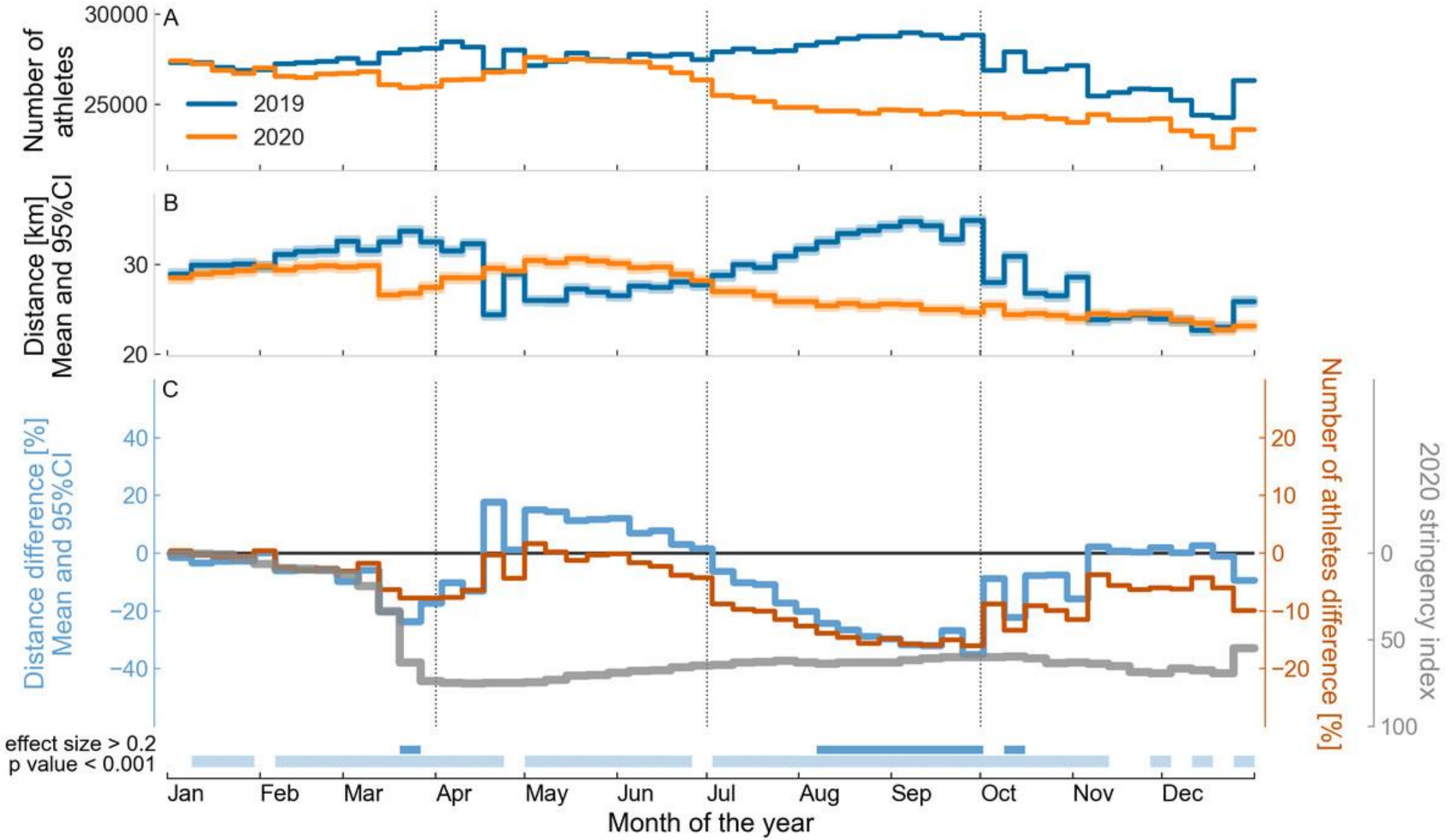
Números gerais 2019 x 2020

Valores/semana	2019	2020	Diferença [%]	Estatística: d, p
Número de atletas	27404 (27103; 27683)	25612 (25222; 25970)	-6,5 (-7,9; -5,2)	1,2; <0,001
Número de corridas	2,46 (2,45; 2,48)	2,41 (2,39; 2,43)	-2,2 (-2,6; -1,7)	0,05; <0,001
Distância [km]	29,2 (29,0; 29,5)	27,0 (26,8; 27,3)	-7.5 (-8,0; -7,0)	0,16; <0,001
Duração [min]	160,5 (159,3; 161,8)	149,7 (148,4; 151,1)	-6.7 (-7,2; -6,2)	0,14; <0,001
Pace [min/km]	5,85 (5,79; 5,95)	5,73 (5,65; 5,87)	-2.0 (-4,4; 0,7)	0,01; 0,116

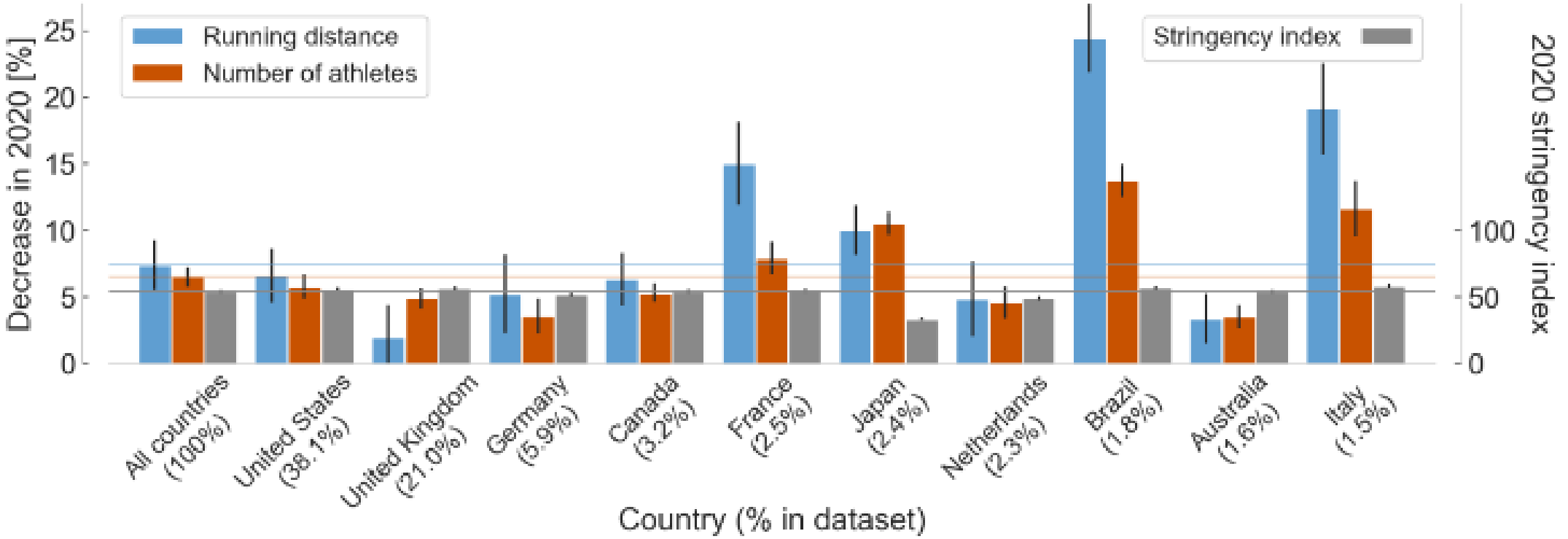


UFABC

Efeitos associados do COVID-19 na prática de corrida de longa distância - Mundo

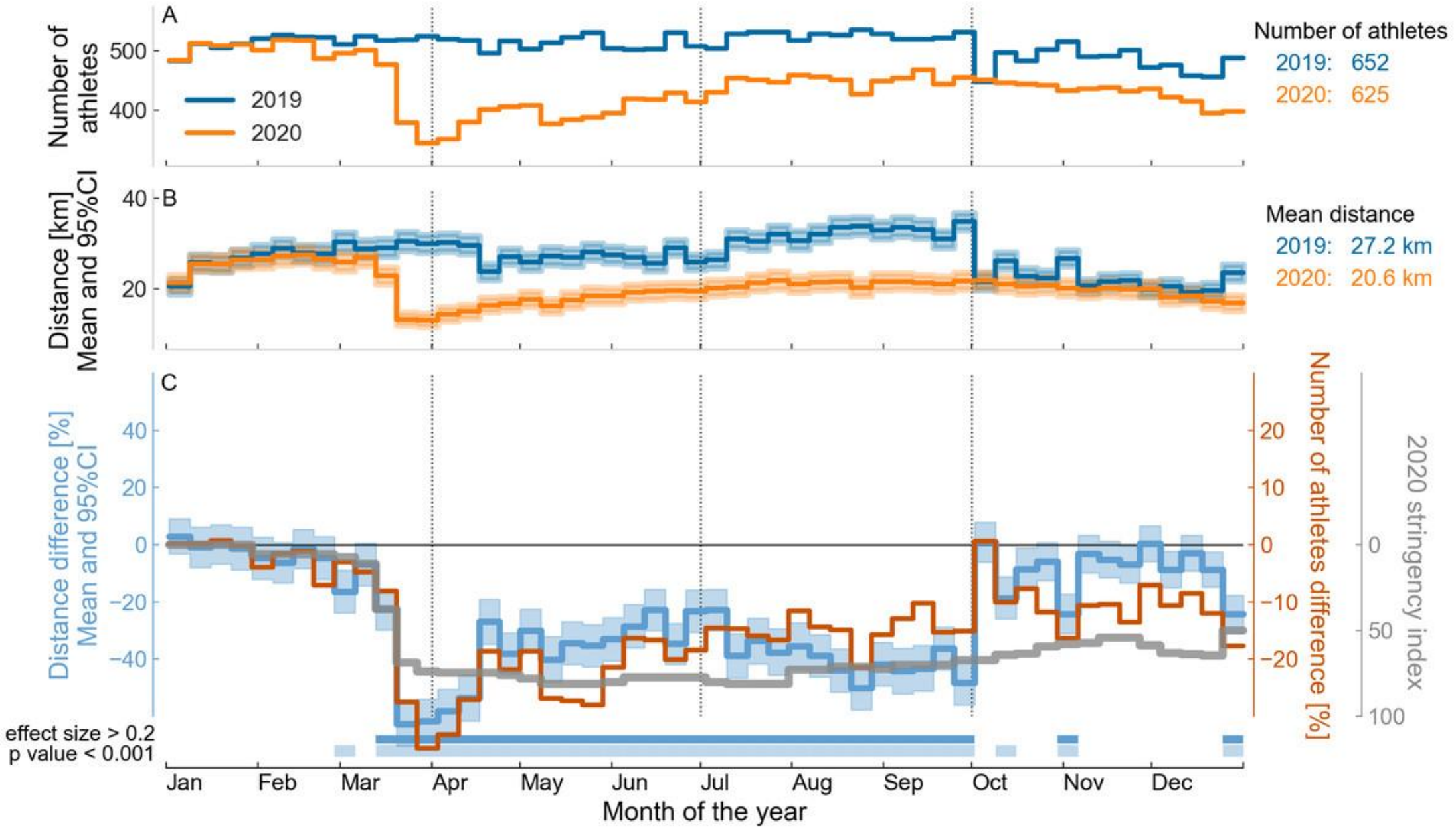


Efeitos associados do COVID-19 na prática de corrida de longa distância - Mundo



UFABC

Efeitos associados do COVID-19 na prática de corrida de longa distância - Brasil



Efeitos associados do COVID-19 na prática de corrida de longa distância - Mundo

Volume de treinamento foi afetado?

⇩ −7,5%

- Mesmos atletas sendo observados ao longo de 2 anos
- Lesões poderiam levar um decréscimo no volume de treinamento | Cancelamento de todas as Majors Maratonas poderia levar uma diminuição de treinos
- Comportamento da queda de treinamento em 2020 **coincide com o início das restrições dos governos**
- Apesar dos bloqueios permanecerem em 2020, é possível observar aumento no segundo semestre quando comparado ao mesmo período em 2019. Seria adaptação? (correr sozinho, usar máscara, horários alternativos)

STRAVA

Aumento no número de atletas

⇧ +40%

Volume de atividades

⇧ +14,7%

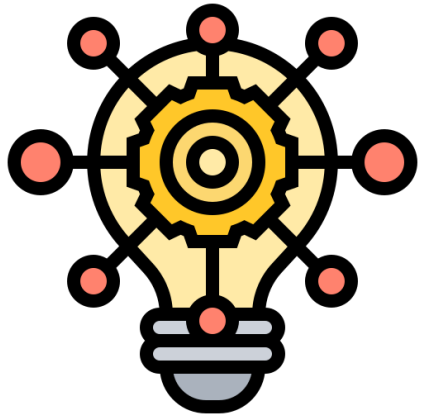
2 estudos encontraram um aumento no volume de corrida durante a pandemia

UFABC

Conclusões

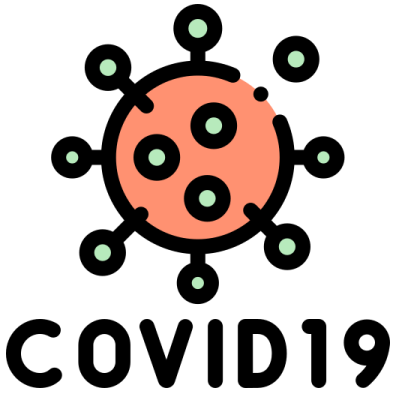
Covid

- Grandes variações no volume de corrida e no número de atletas ao longo de 2020. Possivelmente associados as medidas restritivas



Novos estudos

- Questões específicas a treinamento físico, performance, período pré e pós-competitivo
- Segmentação dos atletas por desempenho



Dataset

- Dataset e notebooks Jupyter disponíveis, permitindo o leitor a reproduzir e ampliar a investigação e gerar novos insights

UFABC

OBRIGADO


BMClab – Laboratório de Biome
bmclab.pesquisa.ufabc.edu.br/pt/
BMClab
Laboratório de Biomecânica e Controle Motor
BMClab
Pessoal
Ensino
Pesquisa
Serviços
Recursos
Publicações
BLOG
Benvindo ao site do BMClab, o Laboratório de Biomecânica e Controle Motor do programa de Engenharia Biomédica da Universidade Federal do ABC.
O BMClab é um laboratório de pesquisa interessado em Biomecânica e Controle Motor do movimento humano, em particular na locomoção e postura humana. Basicamente, Biomecânica estuda a estrutura e função dos sistemas biológicos utilizando o conhecimento e métodos da Mecânica e Controle Motor estuda como os sistemas biológicos controlam seus movimentos. Em um sentido amplo, estamos interessados em saber como seres vivos controlam e executam seus movimentos.
Nós trabalhamos para melhorar a qualidade de vida na sociedade, oferecendo serviços de avaliação em nosso laboratório e na difusão do conhecimento científico.
Próximas disciplinas de graduação ou pós sobre Biomecânica e Controle Motor @ UFABC
Biomecânica I (início em setembro de 2022)
Pesquisas @ UFABC
Efeito da medicação dopaminérgica e dos estágios da doença na marcha de pacientes com doença de Parkinson
Análise clínica da marcha em amputados transtibiais e transfemorais.
Processo seletivo para o mestrado em Engenharia Biomédica
ENGENHARIA BIOMÉDICA
UFABC
Português
English
Busca
Busca
• BMClab sitemap
• Links úteis
Eventos @ UFABC
• BMClab agenda
• Seminários em Engenharia Biomédica
• Seminários em Neurociência
• Todos os eventos
OPEN DATA
BMClab datasets
Leia mais
YES WE CODE
Computação científica @ BMClab
Leia mais