# Závěrečná zpráva z výzkumné stáže v zahraničí

Na začátku letošního akademického roku jsem se vypravila na stáž po boku vědců z renomované britské univerzity University College London. V průběhu této stáže jsem se zapojila do projektu digitalizace záznamů nejstaršího "public-health" programu na světě. V rámci tohoto projektu jsme využívali metody strojového učení a optického rozpoznávání znaků ve spolupráci s renomovanými organizacemi jako Public Health England, Department of Health, WellcomeCollection a Wellcome Trust. Tento projekt byl realizován ve spolupráci s vědci z oddělení ekonomických studií na University College London. Health Visiting Program představuje první program svého druhu v oblasti ochrany veřejného zdraví, jehož počátky se datují do roku 1862.

Kořeny tohoto programu lze nalézt na severu Anglie, konkrétně v Manchesteru, kde ženy ze společensky nižších vrstev navštěvovaly domácnosti žen, které se chystaly na porod. Jejich úkolem bylo poskytovat výuku a rady ohledně přípravy na porod a následné péče o novorozence. Tato služba se výrazně odlišovala od ošetřovatelství a léčby nemocných, a byla prvním příkladem zdravotní sestry, která poskytovala rady a pomoc, nikoli pouze ošetřovatelskou péči. I když se tento program rozšířil po celém území Velké Británie během přelomu 20. století, neexistuje žádná studie datující se do období let 1901 až 1972, která by kvantitativně zhodnotila tento projekt.

Veškeré záznamy týkající se informací o počtu těchto zdravotních sester, jejich pracovním nasazení a provedených návštěvách byly shromažďovány jednotlivými lokálními samosprávami v reportech známých jako "Medical Officer of Health reports". Vzhledem k tomu, že v uvedeném časovém období existuje více než 5800 těchto reportů obsahujících přes 2 miliony tabulek, ruční extrakce dat není ekonomicky efektivní. Další výzvou je, že struktura těchto zpráv nemá žádnou jednotnou podobu, protože každá samospráva měla relativní volnost při jejich vytváření.

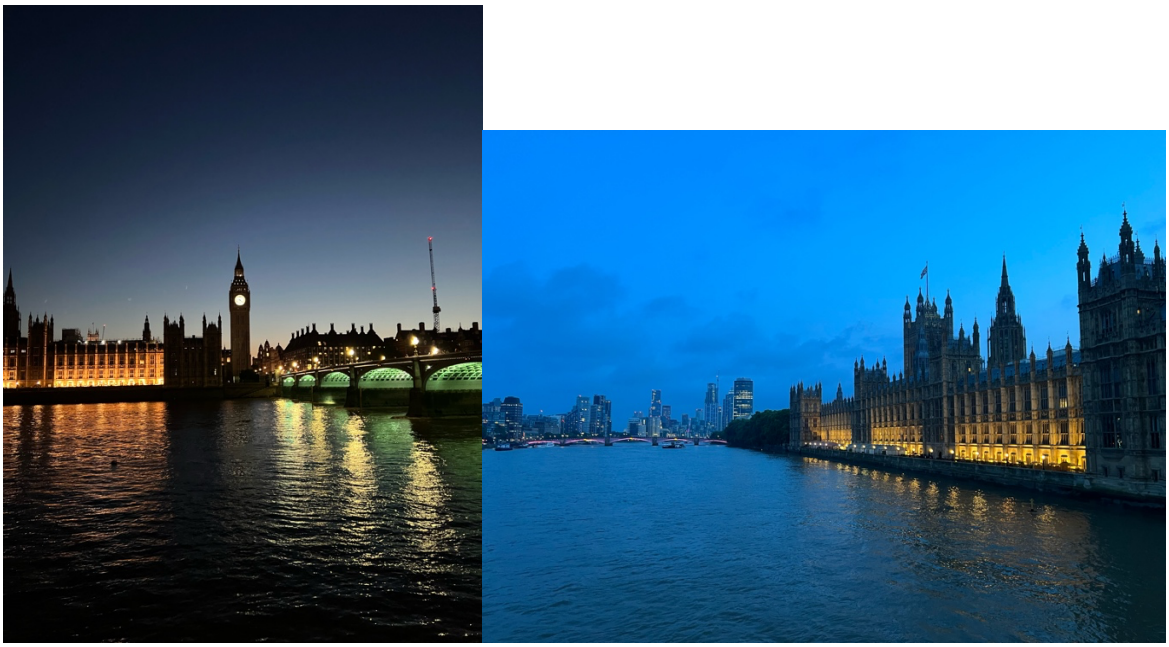
První polovina mé stáže se soustředila na digitalizaci těchto zpráv pomocí technik optického rozpoznávání znaků (OCR) a následná ekonometrická analýza. OCR technologie umožňuje převod tištěných nebo psaných textů na strojově čitelný formát. V případě těchto reportů jsme použili OCR k extrakci textových informací z jednotlivých tabulek a dokumentů. Tím se dosáhlo digitalizace obsahu a umožnilo se snadné vyhledávání, indexování a analýza dat. Přestože struktura zpráv může být různorodá, moderní OCR technologie jsou schopny zvládnout i složité formáty a rozpoznat text správně. Proces digitalizace vyžadoval přizpůsobení OCR algoritmů na konkrétní typy dokumentů a provedení korektur chyb, aby byla zajištěna přesnost převodu textu.

Druhá polovina mé stáže tkvěla v ekonomické analýze dat, které jsme digitalizovali. Ekonomické data jsou ve většině případů volatilní a podléhají změnám vyvolaným vnějšími faktory. V důsledku toho často dochází k tzv. vynechání proměnných a zkreslení výsledků. I když můžeme tyto efekty zmírnit pomocí fixních efektů, často mají tlumící účinek na celý model. Extrahování dat, která skutečně umožňují statisticky významné závěry, ať už se jedná o jejich velikost, rozložení, nebo minimalizaci možných vynechaných proměnných, by mělo být ve středu pozornosti při získávání dat. I když nemáme vždy možnost získat "ideální" druh dat, uznání jejich omezení a důsledků je základním kamenem pečlivé ekonomické analýzy.

Během druhé části mé stáže jsem dospěla k závěru, že často neexistuje vhodný test, nebo může existovat několik testů, které by potenciálně mohly vést ke konfliktním výsledkům. Při řešení obtíží, jako jsou endogenní proměnné, autokorelace, chyba s nesprávným rozdělením, heteroskedasticita nebo multikolinearita, často nalezneme způsoby, jak tyto efekty ovládat nebo je obejít v testovacích metodách a v případě konfliktních výsledků je vždy možné je uvést v našich závěrech. Nicméně, ať už jsou výsledky významné nebo ne, vždy bychom měli usilovat o kritiku jejich hodnoty. Náš proces analýzy dat není bezchybný a téměř na každém stupni trpí našimi předpoklady, zkresleními a určitým nedostatkem náhodnosti. Některé z těchto problémů jsou přítomny i v rámci ekonomické analýzy dat jako celku, například předpoklad, že efekty v ekonomických datech lze zachytit lineárními modely.

*Pohled na Seven Sisters*


*Koncertní síň v londýnském Southbank centre*



*Noční Alžbětina věž, Temže a Houses of Parliament*

Každému, který zvažuje vycestovat do zahraniční na stáž bych vyjet vřele doporučila. Londýn jakožto metropole Evropy umožňuje každému si najít to, co ho baví. Ať už člověka zajímají galerie, sportovní aktivity, divadlo či přednášky a odborné semináře. V neposlední řadě bych tímto chtěla velice poděkovat Fakultě, která mne v rámci Mobilita-Akce 200 na stáži podpořila.