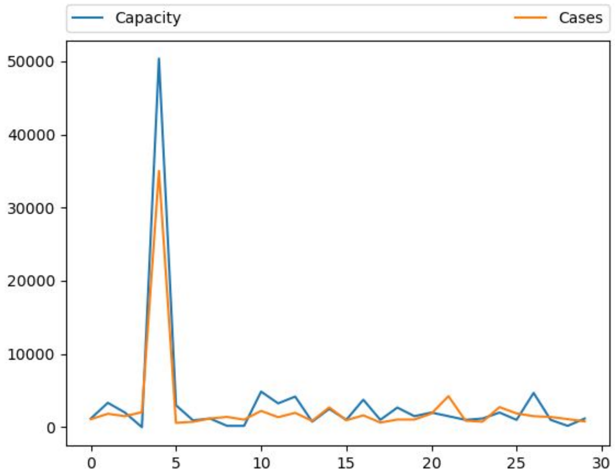
Swabs2Labs

Team Mallocators, IIIT Hyderabad

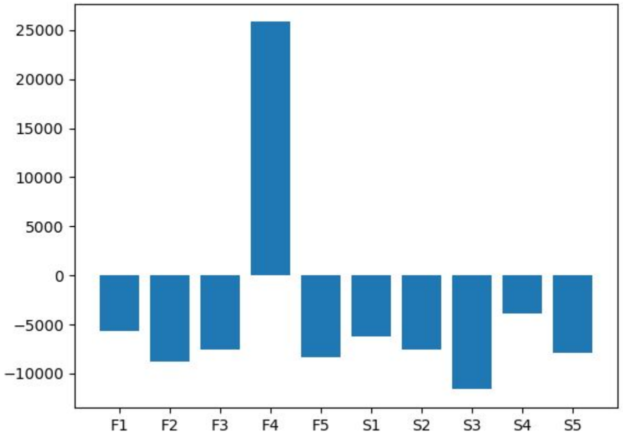
Arpan Dasgupta, Kunal Jain, Nikhil Chandak, Shashwat Goel

Data Exploration

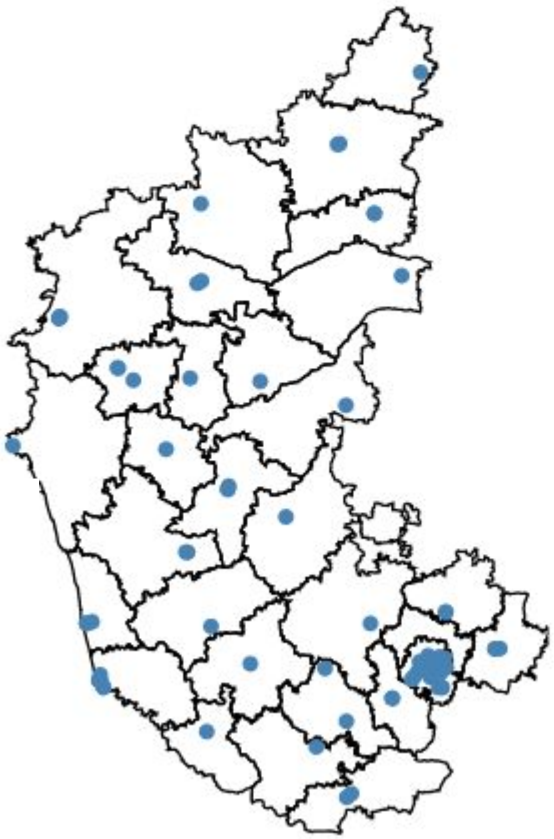
- Finding pairs of labs within 40 km of each other
- Comparing cases and capacity per district
- Bangalore Urban district has a cluster of 34 labs, all within 40km of each other. Other labs are sparsely distributed so small clusters.
- Capacity deficit on samples



Comparison of capacity vs cases in each district



∑Capacity - ∑Samples on different dataset



Geographic distribution of labs

Problem Interpretation

- There are 30 districts in the state of Karnataka. Each district accumulates all it's swabs (samples) at a headquarters (HQ), the coordinates of which are provided.

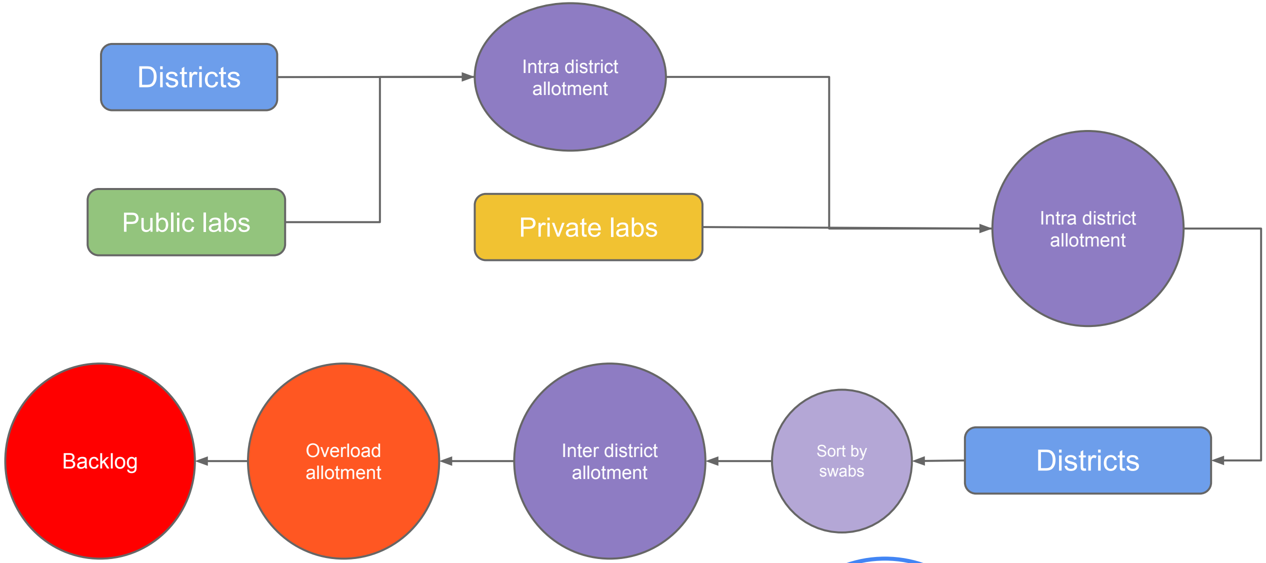
- There are 86 labs, one of two types: private (cost per sample: 1600) or public (cost per sample: 800). Each lab has a fixed per-day capacity, above which only the district where the lab exists can overload upto 100 samples (with a penalty of 5000 per sample overloaded).

- The samples are distributed to 'labs' within the district or elsewhere. With a high penalty (10,000), samples can be kept as backlog at the HQ to be processed the next day. We thus optimize on a day-to-day basis.

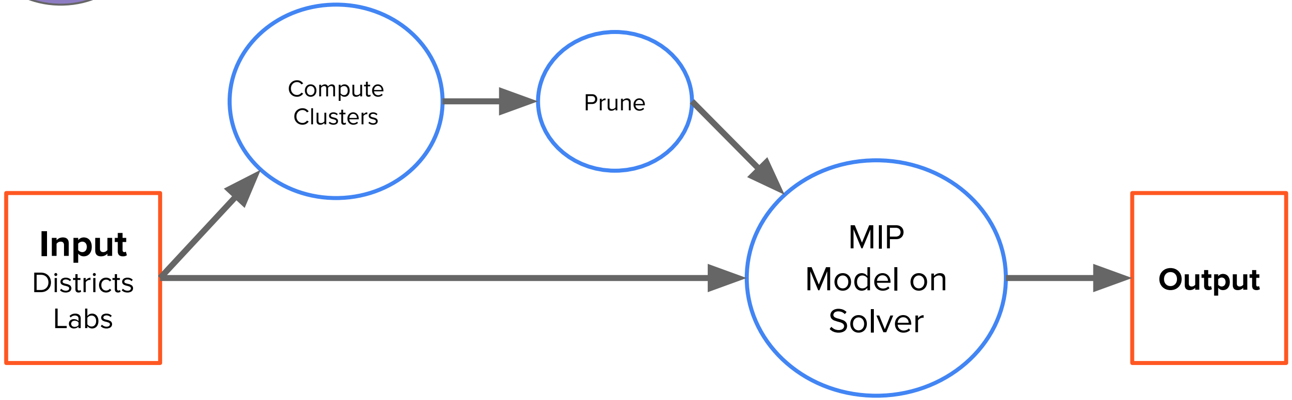
- When transferring samples to another district, the euclidean distance to a 'centroid' of a chosen ‘cluster’ of labs (provided any pair of labs in the cluster is within 40km of each other) can be taken as a fair approximation of transport costs (when multiplied by 1000 per km).

Approaches

Baseline Algorithm



Mixed Integer Programming



MIP Formulation - Intra District Model

- First we formulate the MIP model for a simplified version of the problem where we assume no outside district transfers (to external labs) are allowed. This gives us the initial model -

d_i is the number of samples at district HQ i

cap_j is the capacity of lab j

x_{ij} represents the amount of samples district i sends to an internal lab j

o_j represents how many samples are being overloaded in lab j

$$\min_{x,o} \left(\sum_i \sum_j X_j \cdot x_{ij} + \sum_j 5000 \cdot o_j + \sum_i 10000 \cdot (d_i - \sum_j x_{ij}) \right)$$

$$\sum_j x_{ij} \leq d_i \qquad \forall\, i \qquad (1)$$

$$\sum_i x_{ij} \leq cap_j + 100 \qquad \forall\, j \qquad (2)$$

$$\sum_i x_{i,j} - cap_j \leq o_j \qquad \forall\, j \qquad (3)$$

$$x_{ij}, o_j \in \mathbb{Z}_{\geq 0}$$

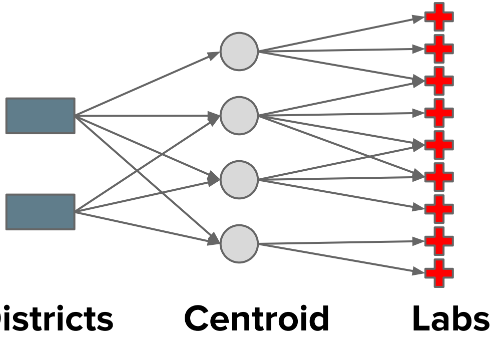
$$X_j \in \{800, 1600\} \text{ depending upon the type of lab } j$$

MIP Formulation - Introducing Clusters

- The model till now doesn't handle transfer to external labs. Interpret the concept of centroid as an intermediate center between districts and outside labs:
  - Facilitate distribution of incoming samples among the cluster's labs
  - Also saves transportation cost.

- Backward Formulation:
  - First prepare clusters
  - Afterwards choose which external labs should a district send its samples through a centroid.

- Introduction of new variables in the formulation:
  - To model samples sent from district to centroid
  - To model samples received by labs from centroid
  - Precise details and derivation are available in the Report.



Mixed Integer Programming - Final Model

p_{ic} represents the number of samples received by the centroid c from district i.

q_{cj} represents the number of samples going from centroid c to lab j

z_{ij} is a binary variable representing whether district i is sending any samples to centroid j

And, ƛ = very large number

$$min\left(\sum_j X_j\cdot\left(\sum_i x_{ij}+\sum_k q_{kj}\right)+\sum_j 5000\cdot o_j+\sum_{i,c}1000\cdot dist(i,c)\cdot z_{ic}+\sum_i 10000\cdot\left(d_i-\sum_j x_{ij}-\sum_c p_{ic}\right)\right)$$

[Distance constraint]	$\sum_j x_{ij}+\sum_c p_{ic}\leq d_i$	$\forall\ i$	(1)
[Inflow = Outflow on centroid c]	$\sum_j q_{cj}=\sum_i p_{ic}$	$\forall\ c$	(2)
[External inflow to lab <= Capacity]	$\sum_k q_{kj}\leq cap_j$	$\forall\ j$	(3)
[Total to lab <= Cap. + 100 (overload)]	$\sum_i x_{ij}+\sum_k q_{kj}\leq cap_j+100$	$\forall\ j$	(4)
[Overload >= Total - capacity]	$\sum_i x_{ij}+\sum_k q_{kj}-cap_j\leq o_j$	$\forall\ j$	(5)
[Set z if district-> centroid transfer]	$\lambda\cdot z_{i,j}\geq p_{i,j}$	$\forall\ (i,j)$	(6)
[Max. one centroid per district]	$\sum_j z_{ij}\leq 1$	$\forall\ i$	(7)
[Each lab of the cluster should receive at least as many sample as districts sending to its centroid]	$q_{j,k}\geq\sum_i z_{i,j}$	$\forall\ (j,k)$	(8)

$$x,p,q,z,o\in\mathbb{Z}_{\geq 0}$$

$$X_j\in\{800,1600\}\ \text{ depending upon the type of lab } j$$

Cluster Generation & Selection

Graph Formulation

Vertices = Labs. Edges between labs <= 40km apart.

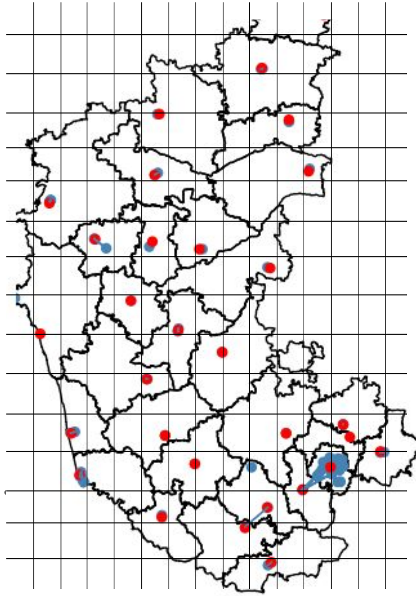
Clusters are now equivalent to cliques in this graph.

Find clusters of size <= k using backtracking search.

Spatial Model

Divide map into grid of R*C cells (keeping one cell within 10km^2 area)

With each intersection as center, pick labs within 20km radius as a cluster.



Combining: DFS to compute connected components, backtracking in each.

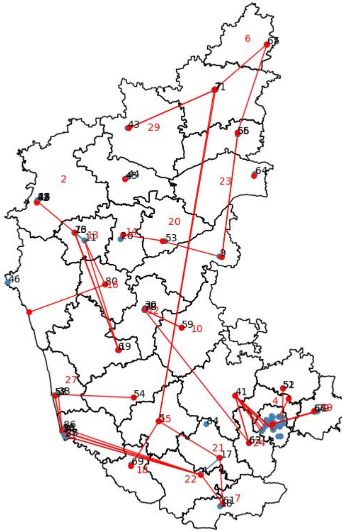
Selecting Clusters for Input

Pick high capacity clusters, and then randomly pick them to cover all labs.

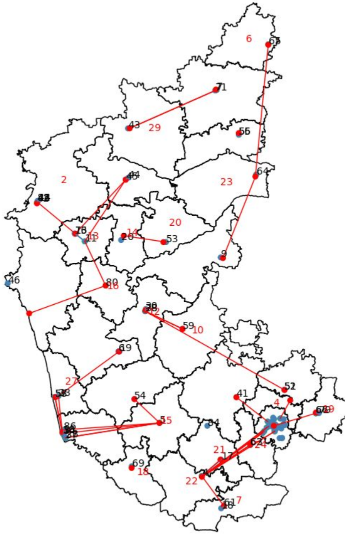
- Must give a clique cover of labs
- A single lab shouldn't come in too many cliques.

Testing and Results

- Calculating lower bounds and margin of closeness
- Benchmarked all approaches against the sample outputs
- Performance testing on varying case load
- Output comparisons on short and long runs of MIP solver in Python 3.6 using PythonMIP library.



Baseline	MIP (10mins)	MIP (30 mins)	Best Score	Closeness
138,782,153	121,942,529	121,779,813	121,403,144	99.807%
162,357,199	142,503,116	142,459,632	142,407,282	99.817%
151,810,345	131,316,815	130,934,769	130,814,729	99.853%
124,178,874	110,211,626	110,196,392	110,180,913	99.863%
157,921,190	139,749,409	138,749,708	138,607,286	99.878%



Lower bound

- A tight lower bound gives a realistic idea of the optimization landscape
- We know: Our MIP model provides optimal solutions provided the right clusters

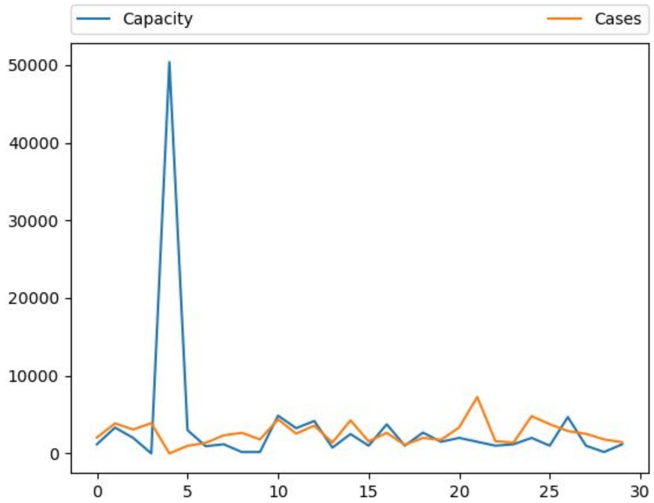
- Can we upper-bound the error due to suboptimal cluster input?
  a) Relax constraint (8): ≥1 sample to every lab in cluster
  b) We need: $\forall$ valid combinations of labs (clusters) **L**, $\exists$ a clique **C** s.t. **L** $\subseteq$ **C**
  c) So we just take all maximal cliques! Feasible for Karnataka's lab distribution

- We now get optimal combination of labs, but the centroid input to MIP is inaccurate. Actual centroid is centroid of chosen subset of labs for each district.
- Upper-bound on this error: $1000 * \Sigma R_{cluster(district)}$ over all districts
- Therefore, tight lower bound = $O_{MIP'} - 1000 * \Sigma R_{cluster(district)}$

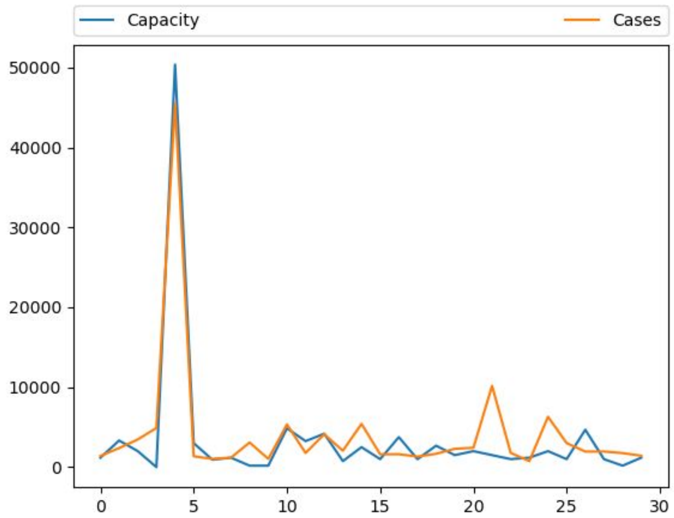
Detailed proof in report!

Robustness

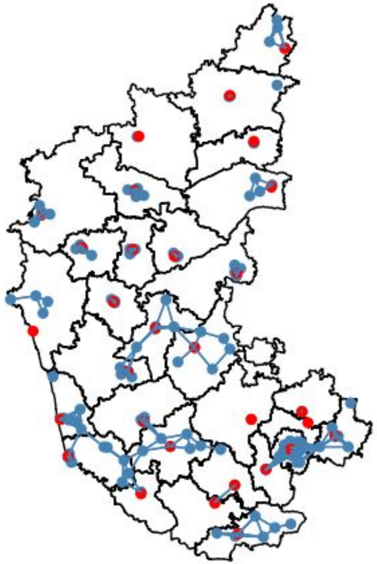
Dataset	MIP (in 10mins)	Lower Bound	Optimality
Bangalore - 0 Rest - 1.8x	112,886,508	112,777,320	0.097%
(1 - 2.5)x load	553,013,565	552,861,237	0.028%
More labs made	425,153,802	423,837,100	0.311%



Drastic change in sample distribution
(Bangalore now has 0 samples)



Samples far exceeding capacity (congestion)


Much more labs (155 in above!)

Thank You

The challenge has a tangible, real world impact. We thoroughly enjoyed developing fast near-optimal approaches to this pressing problem while also gaining experience in Operations Research.