

The methodology used for estimating m and u probabilities:

The model settings were as defined in the full configuration file at the time of training:

https://github.com/ministryofjustice/hmpps-person-match/tree/main/hmpps_cpr_splink/cpr_splink/model/model_files

Training proceeded as follows:

```
from splink import DuckDBAPI, Linker, block_on

from hmpps_cpr_splink.cpr_splink.model.settings import settings

deterministic_rules = [
    block_on("name_1_std", "last_name_std", "date_of_birth",
"postcode_arr[1]"),
    block_on("name_1_std", "last_name_std", "date_of_birth",
"postcode_arr[-1]"),
    block_on(
        "name_1_std",
        "last_name_std",
        "sentence_date_arr[1]",
        "postcode_arr[1]",
    ),
    block_on(
        "name_1_std",
        "last_name_std",
        "date_of_birth",
        "sentence_date_arr[-1]",
    ),
    block_on("name_1_std", "pnc_single"),
    block_on("name_1_std", "cro_single"),
    # sometimes we see like 1900/0054321R vs 054321/12U
    """
    l.name_1_std = r.name_1_std
    and l.last_name_std = r.last_name_std
    and substring(l.pnc_single, 7, 6) = substring(r.cro_single, 1, 6)
    """,
    """
    l.name_1_std = r.name_1_std
    and l.last_name_std = r.last_name_std
    and substring(l.cro_single, 1, 6) = substring(r.pnc_single, 7, 6)
    """,
]
recall = 0.8

linker = Linker(cleaned_data_table, settings=settings, db_api=db_api)

linker.training.estimate_probability_two_random_records_match(
    deterministic_rules,
    recall=recall,
)

em_blocks = [
    block_on("pnc_single"),
    block_on("cro_single"),
    block_on("name_1_std", "last_name_std", "date_of_birth"),
```

```
        block_on("postcode_arr[1]", "sentence_date_arr[1]"),
    ]

linker.training.estimate_u_using_random_sampling(max_pairs=1e8)

for blocking_rule in em_blocks:
    linker.training.estimate_parameters_using_expectation_maximisation(
        blocking_rule,
    )
```

The ‘master_defendant_id’ and ‘override_marker’ parameters were set manually separately to the training process because these represent manual overrides and as such are not data-driven estimates.