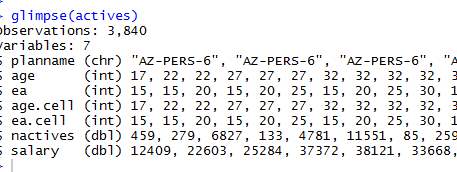
9/30/2015

# Prototype data and methods

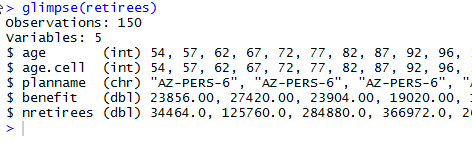
## Data files

4 data files

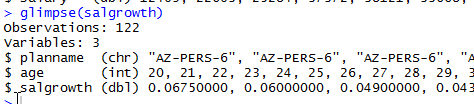
* actives.rda - # actives and average salary (annual, dollars) by age, ea; also gives the original grouped cell an observation fits in, for interpolated data



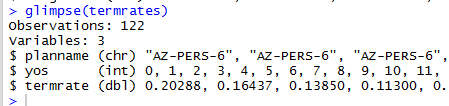
* retirees.rda # retirees and average benefit (annual, dollars) by age; also gives the original grouped cell an observation fits in, for interpolated data



* salgrowth.rda – annual salary growth rate by age 20:80, as a decimal, NO MISSING AGES; all inclusive – inflation + productivity + merit



* termrates.rda – separation (withdrawal) rates by yos 0:60, as a decimal, NO MISSING YOS; all inclusive; does not include mortality or disability decrements



## Plans

There are 3 plans:

* AZ-PERS-6 – “average” plan, based on Arizona Public Employees Retirement System; ppd\_id=6
* LA-CERA-43 – “mature” plan, based on Los Angeles County Employees Retirement System (general employees); ppd\_id=43
* WA-PERS2-119

In the actives and retirees data frames, each plan has grouped data and one or more interpolated versions that spread data across ages and ea.

In interpolated versions, nactives and nretirees are spread uniformly:

* Nactives: uniformly across allowable cells (age>=ea) in the age-ea group
* Nretirees uniformly across the age group (retirees).

In all interpolated versions, salary and benefit are adjusted so that sum of pay and sum of benefits in each group match the values in the original grouped data.

The two interpolated salary versions differ as follows:

* scaled: identified by planname suffix “.fillin” -- salary data are scaled within each original age-ea group so that they follow merit+productivity relationships in the plan’s salary scale (note that the plan’s assumed inflation is excluded)
* uniform: identified by planname suffix “.fillin.unif” -- salary data are identical for each age-ea combination within a given original age-ea group

The benefit values are spread across ages using a spline, and then adjusted within each original age group to ensure that total benefits within each group match total benefits in the original grouped data.

## Plantypes

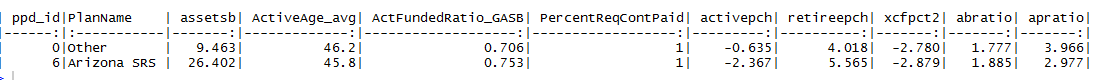
3 plan types

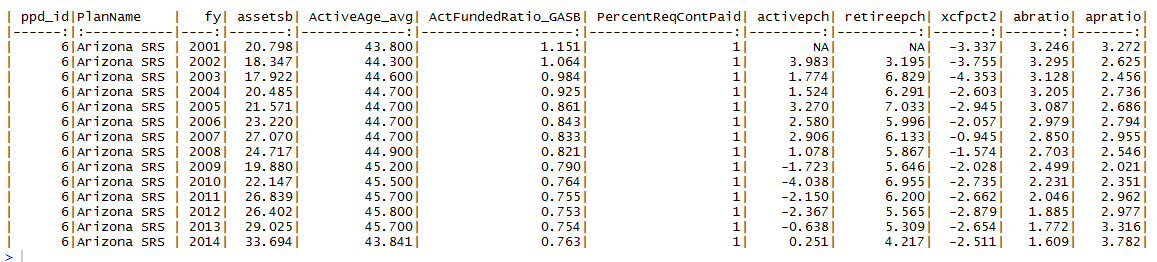
* average
* mature
* immature

All 3 based on examination of PPD data EXCLUDING safety plans

### Average plan

#### Chosen plan AZ-PERS ppdid 6

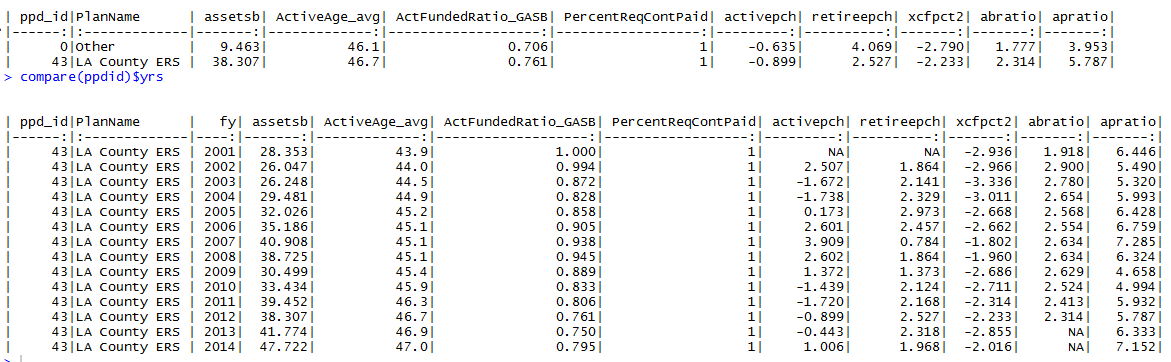




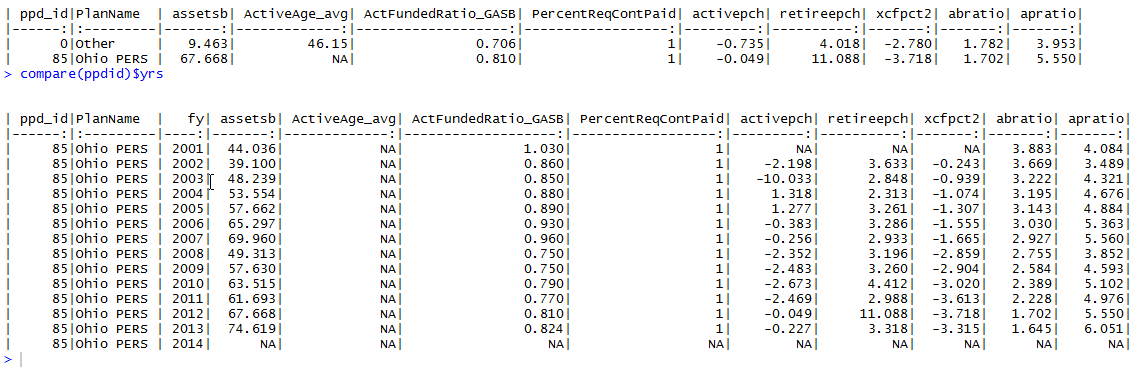
### Mature plan

Am thinking OHIO PERS 85 might be better choice.

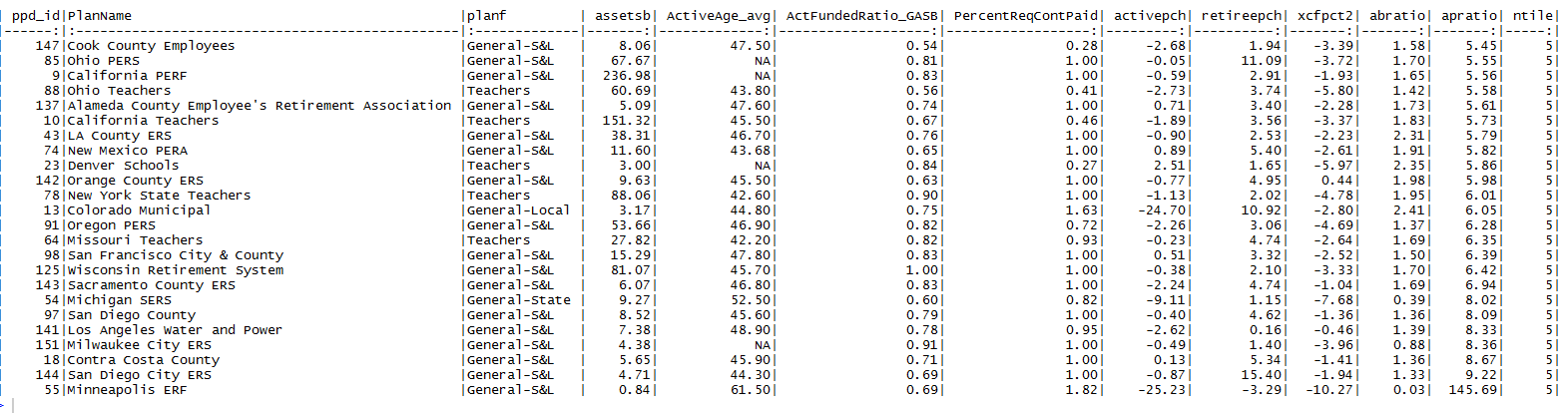
#### Chosen plan LA County ERS ppdid 43



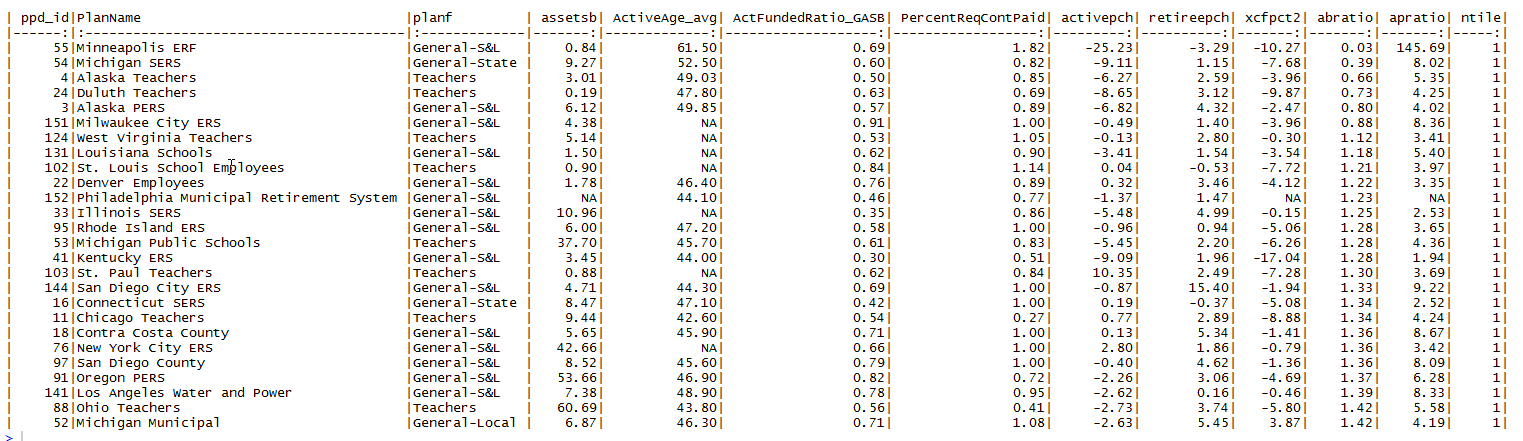
#### Other mature plans



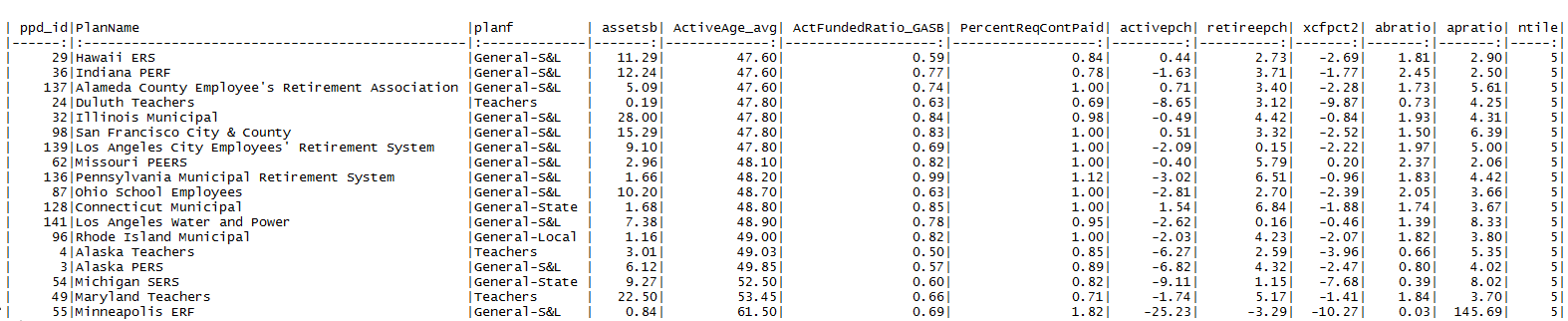
##### High apratio plans



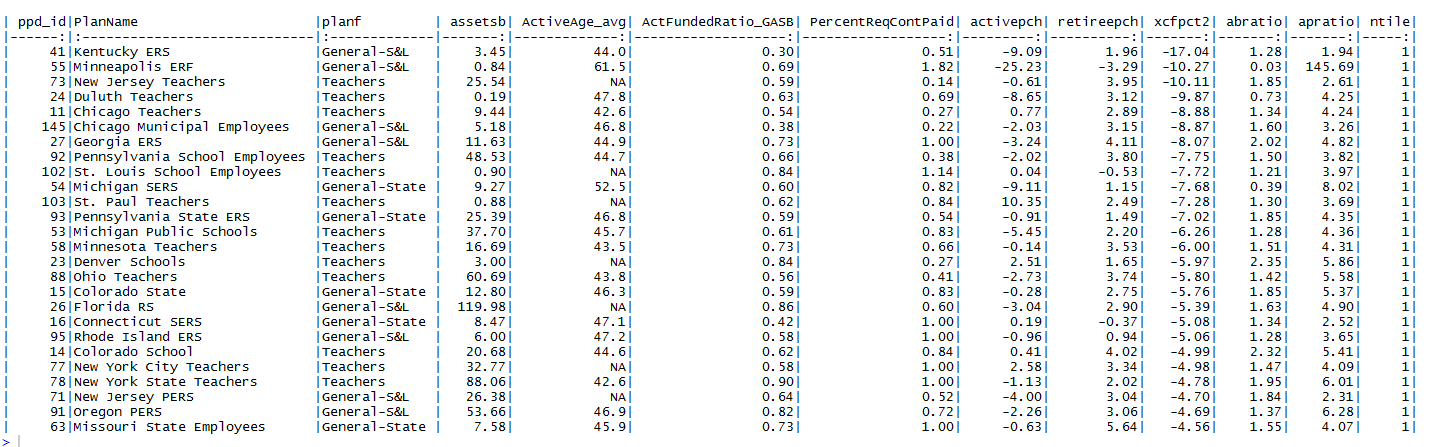
##### Low abratio plans



##### High age plans

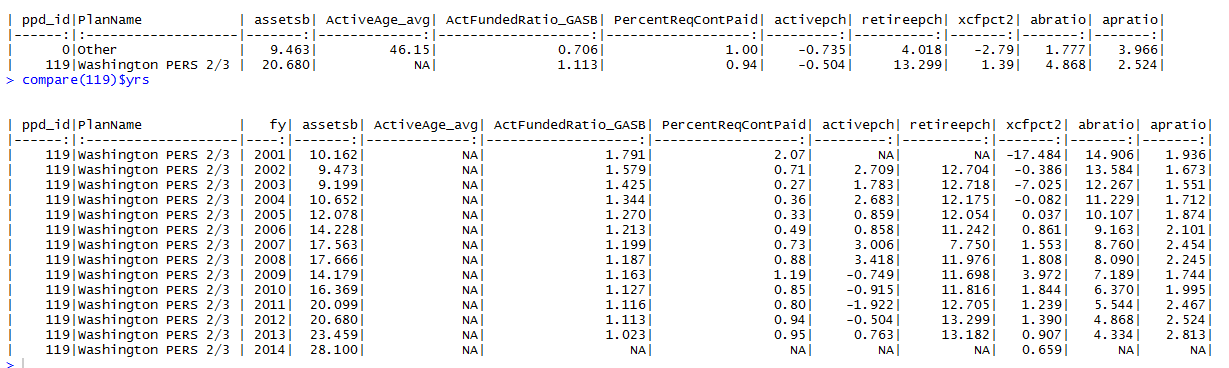


##### High negative xcfpct plans



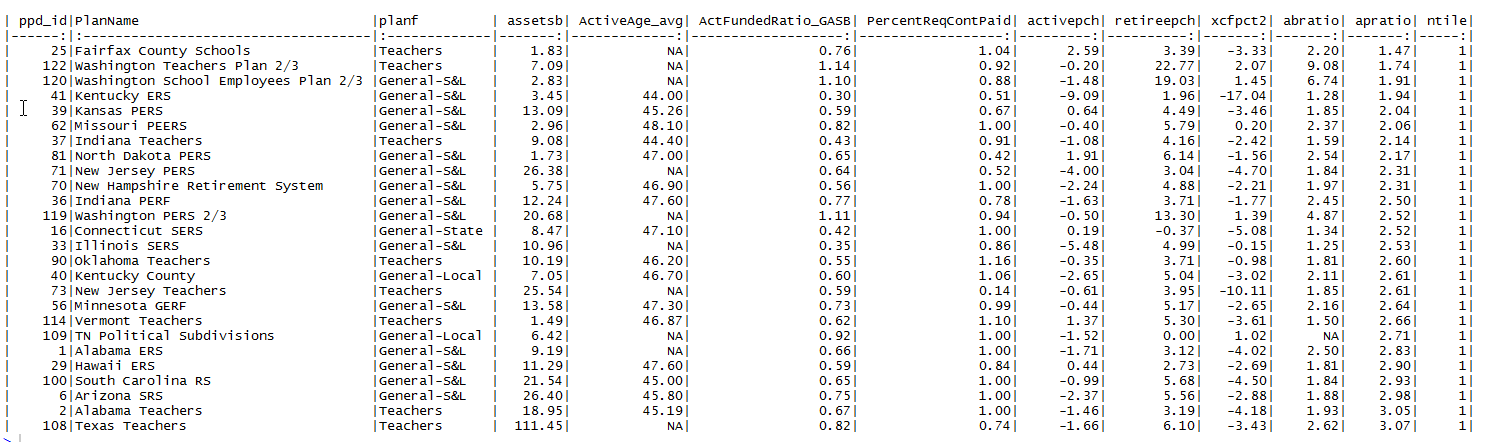
### Immature plan

#### Chosen plan WA PERS 2/3 ppdid 119

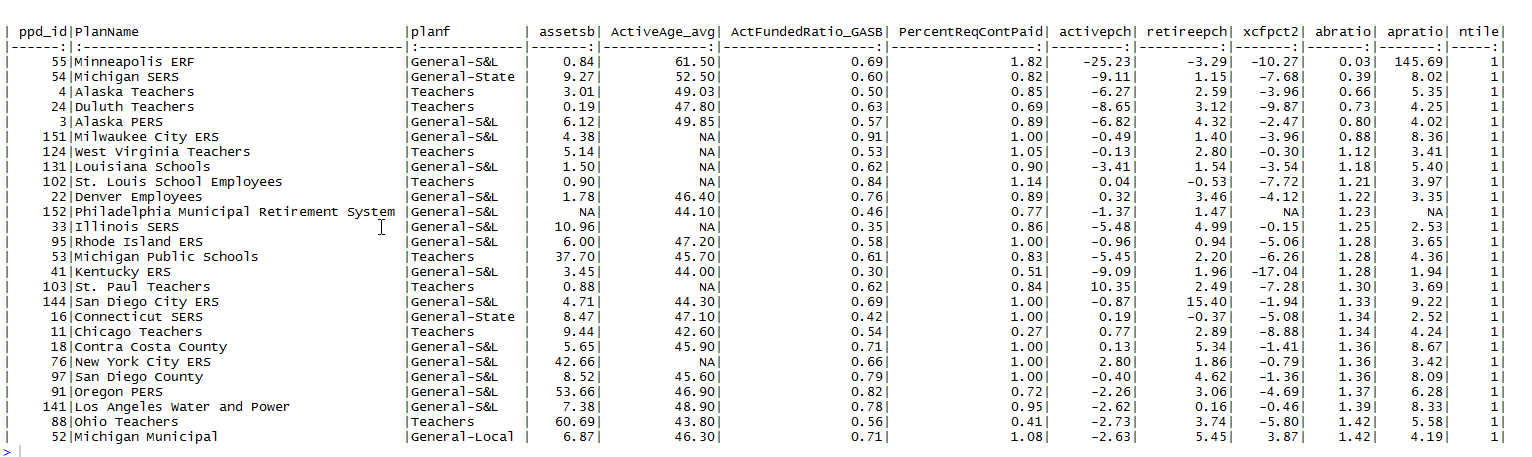


#### Other immature plans

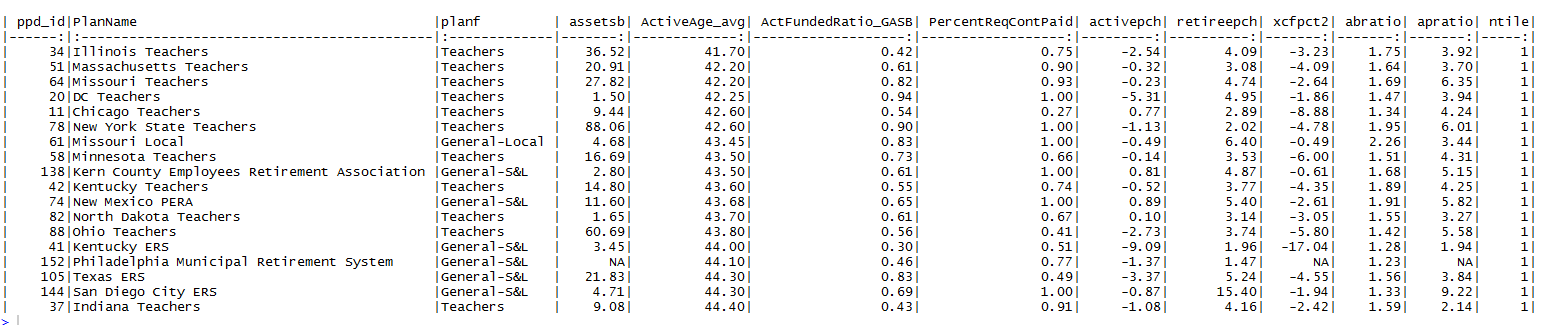
##### Low apratio plans



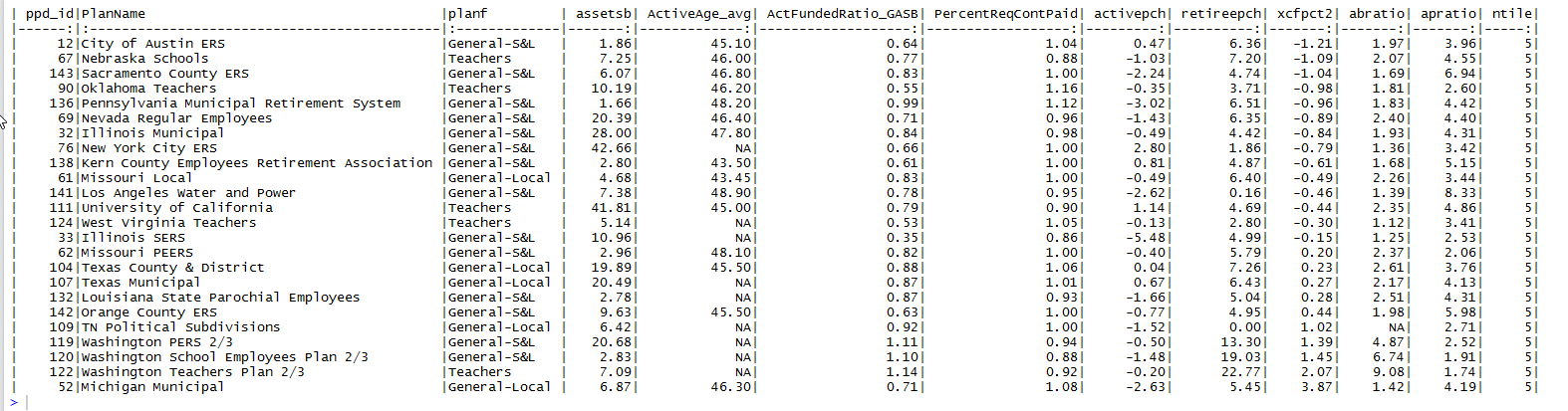
##### High abratio plans



##### Low age plans

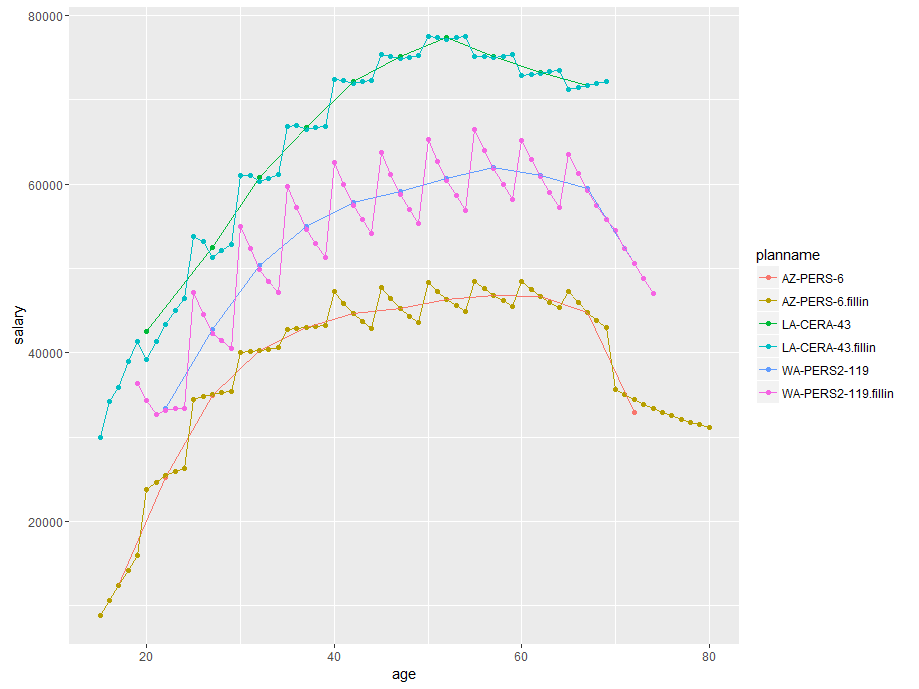


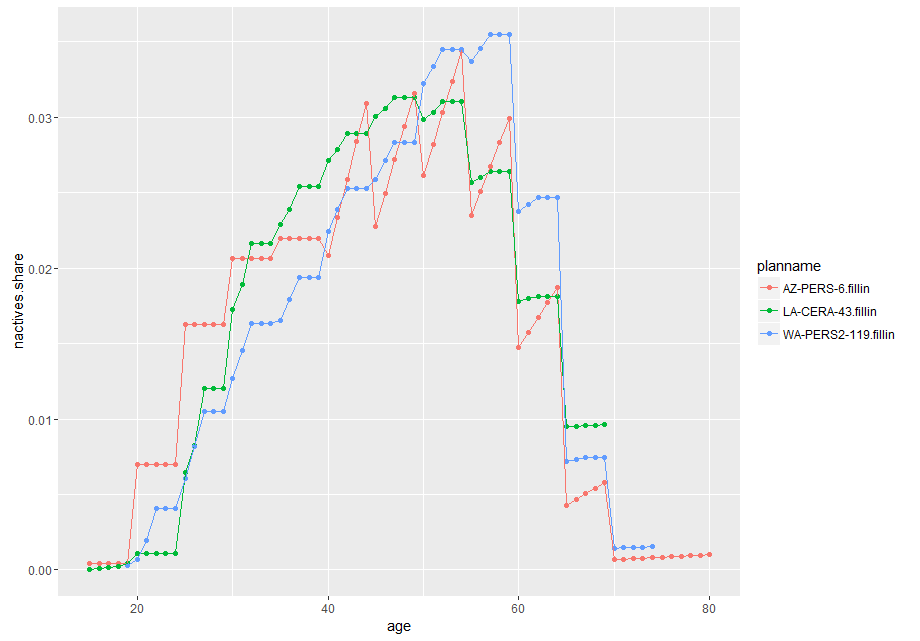
##### Low negative xcfpct plans



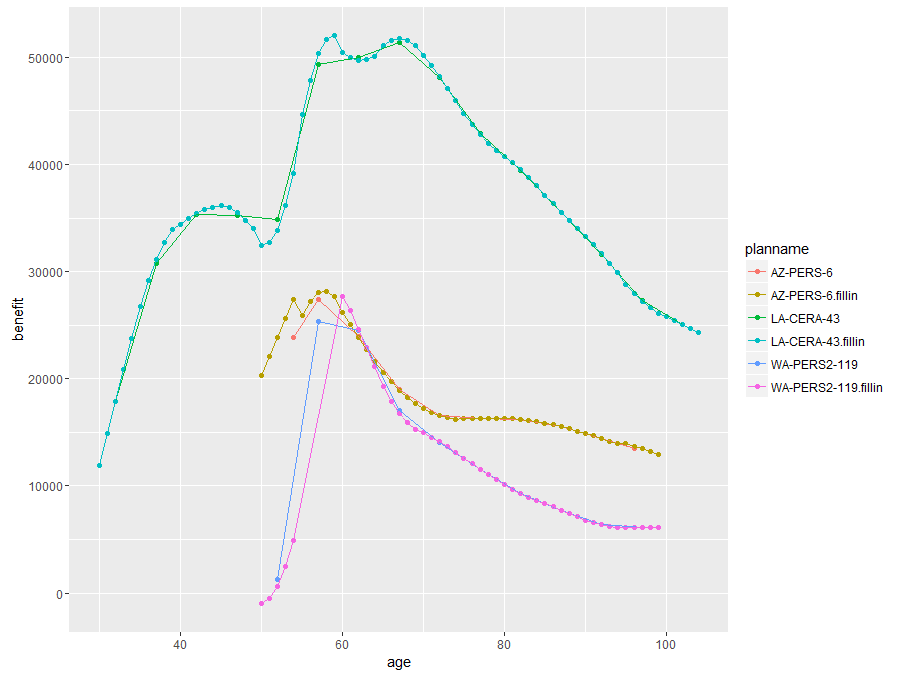
## Results

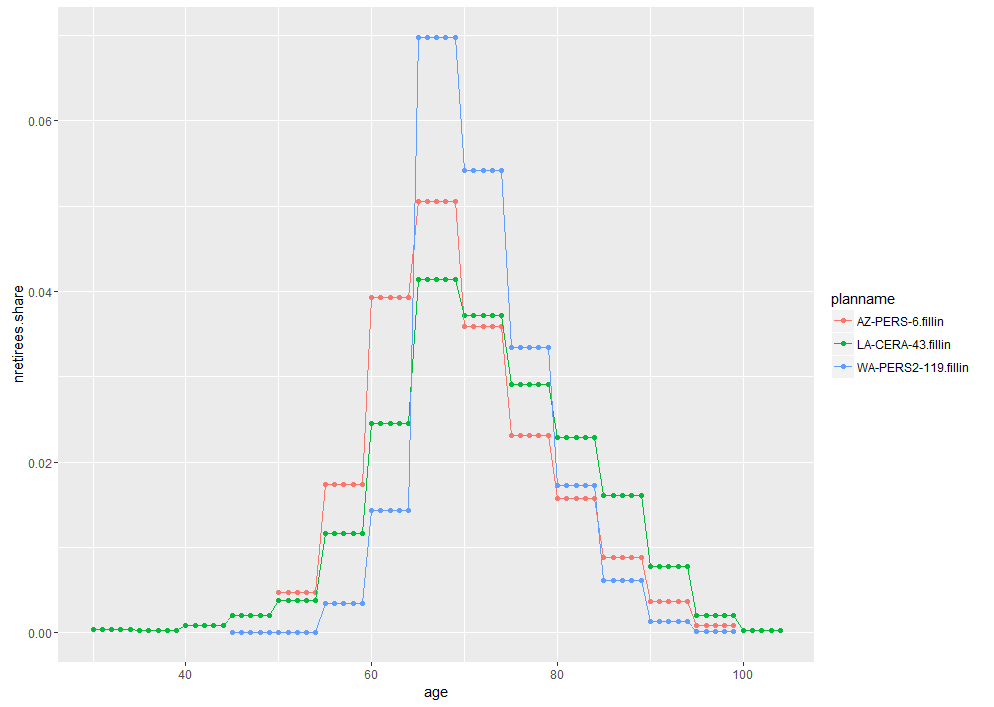
### Actives



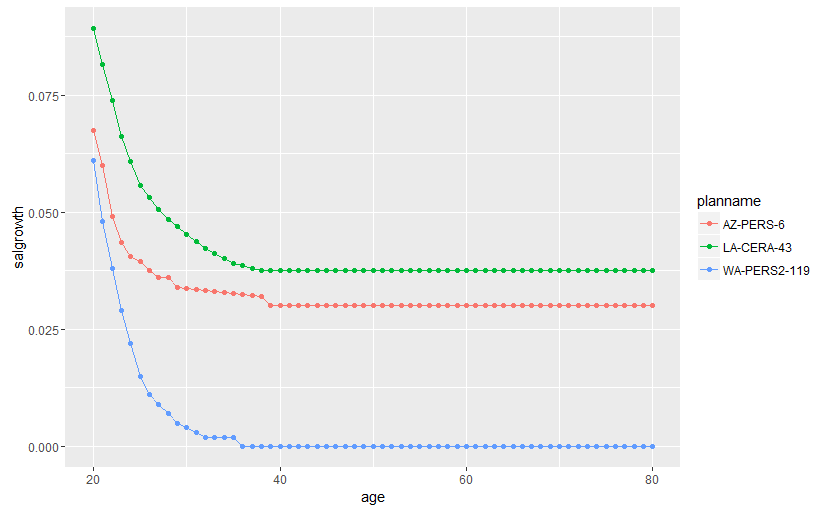


### Retirees



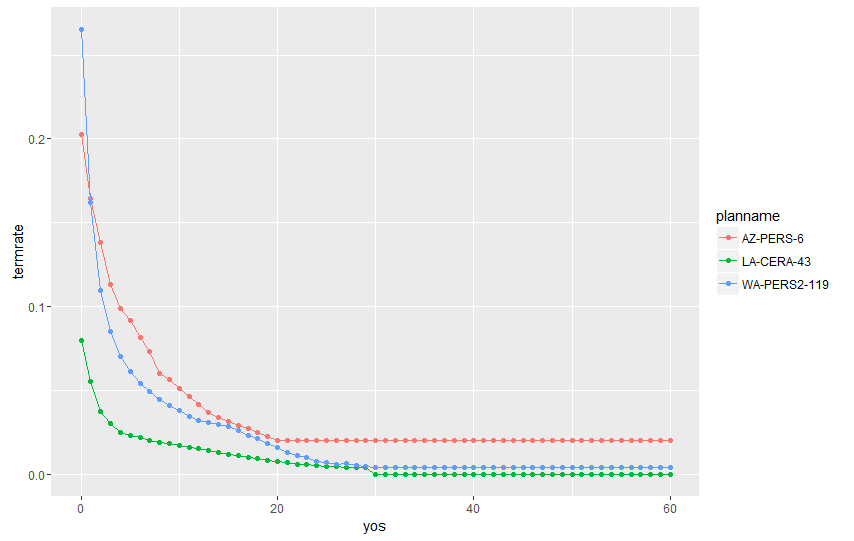


### Salary growth



Hmmm…I have to look at salary adjustment for WA PERS2

### Term rates



# Appendix: Interpolating and extrapolating nactives and avg salary by individual years of age and ea

## Data and goal

The basic data we have to work with are:

* Matrix of nactives, grouped, age x ea (usually formed from age x yos data by subtracting yos from age)
* Matrix of avg salaries, grouped, age x ea (usually formed from age x yos). Note that this is different than what you used for TPAF, where you only had a vector of average salaries by age group, not a matrix of age x ea.
* Salary growth rate schedule by individual ages (sometimes extended from age-group data using spline)

We want to ungroup (spread) these data over the full range of potential ages and ea, to get rid of lumpiness in the grouped data (as when an entire group of workers retires in one year, and there are no new retirees for another 5 years).

The filled-in data are included in the actives data frame with the original planname and the “.filledin” suffix.

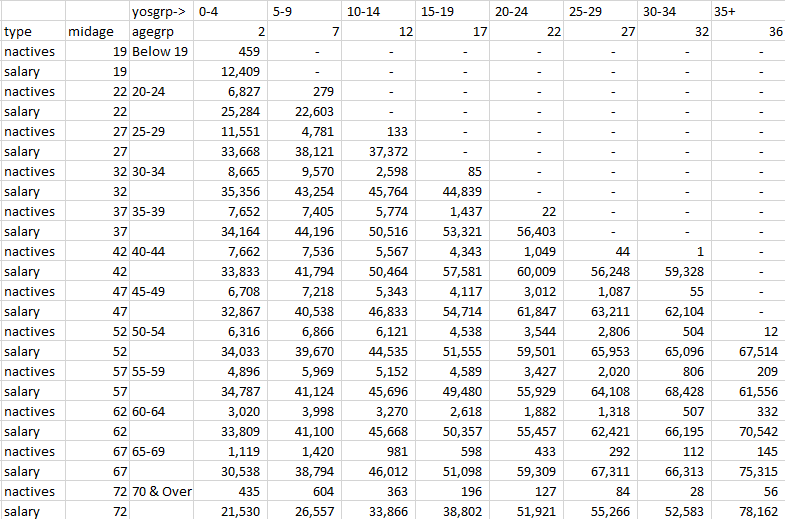
## Method

1. **Determine group sizes:** The data pulled into the model do not show the actual group bounds, just the approximate age and yos midpoints of each group, so I had to assume/estimate the bounds so that I know what ages to spread each group over. (I could change the data to include lower and upper bounds of each group but that may be more trouble than it is worth.) Wanted an automated method of defining the groups so it would work with any plan. The method I chose is (a) first group starts at zero and stops at the first group’s point. (e.g., first age group is 10-19 and first ea group is 10-10 – for the person noted above); (b) last group ends at age 80; (c) in-between groups have width equal to the distance between midpoints, with the proviso that every midpoint must fall into a group.
2. **Spread actives:** Just spread them uniformly across the allowable cells (age>=ea) in their group. We could consider spline methods but I am not sure we really know another distribution we should use.
3. **Spread average salary** in 3 steps:
   1. Use the group average for each allowable age-ea combination in the group
   2. Scale salary for each age-ea combination to the group average salary using the age-based salary scale for the plan. For example if a group has ages 40-44 and an average age of 42, then the salary for all age 44 rows in the cell will be scaled based on the relationship between age 44 salary and age 42 salary in the salary scale
   3. Scale all salaries in the group proportionately so that the total pay in the group equals the total pay in the original grouped data. For example if the original grouped data had total pay of $505k and if the filled-in data total pay of $500k after step “b”, all salaries in the group would be multiplied by 1.01.

## Results for AZ-PERS

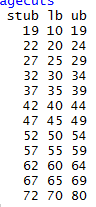
### Raw data

Here is nactives and avg salary data taken from the AV in age x yos format. yosgrp and agegrp are the groups actually shown in the AV. I had to choose a specific age and yos for each group by hand (needed even if we don’t do interpolation). The age I chose for agegrp is midage; the yos I chose is the row below yosgrp. (We could add two rows and two columns to the data to give the bounds, if needed, and skip the step of constructing bounds.) Note that there is 1 person in the 40-44 age group with 30-34 yos;. Based on midpoints we assign this person ea=10 (42 – 32). If there were a lot of these occurrences we might have to start imposing cleaning rules on the data but can live with this I think.

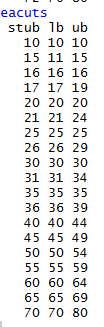


### Constructing the bounds over which to spread each group

Here are the cutpoints chosen by the method. Stub is the midpoint in the grouped data, lb is lower bound, ub is upper bound. Thus, for agecuts stub corresponds to midage above. lb and ub give the range over which the grouped data will be spread.



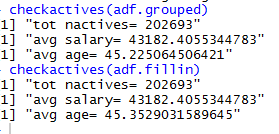
Here are the cuts for ea. It looks a little strange, due to the conversion from yos to ea.



I think this is good enough as a reconstruction method but it is not the same as actual bounds in the data. We could modify the data to include the bounds.

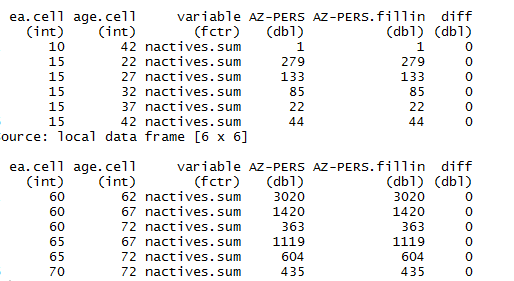
### Grand totals

Grand total # actives and avg salary match exactly. Avg age is slightly higher in the filled-in data. (For what it is worth, per the AV, average age of actives is 45.7 so the move is in the right direction.)

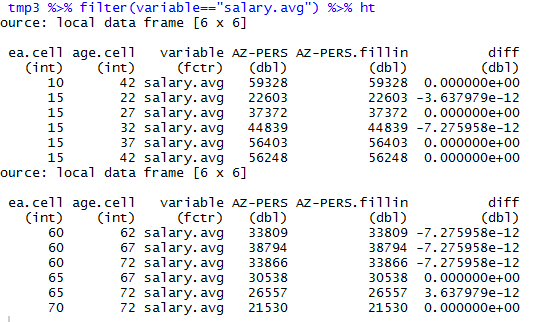


### Aggregates within each grouping

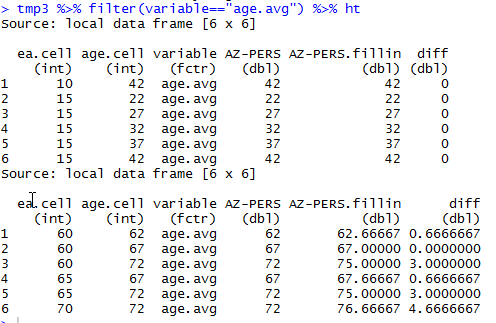
Nactives match



Avg salary matches



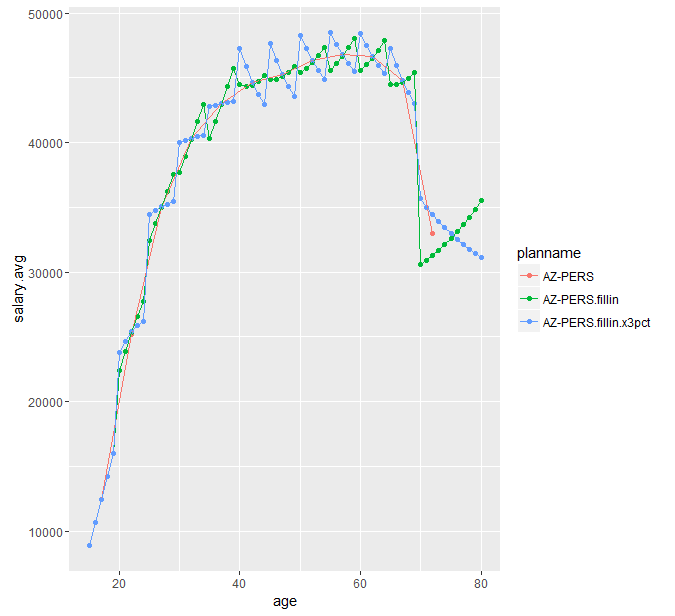
Avg age does not always match but is close except for a few high-age records



### Graphs

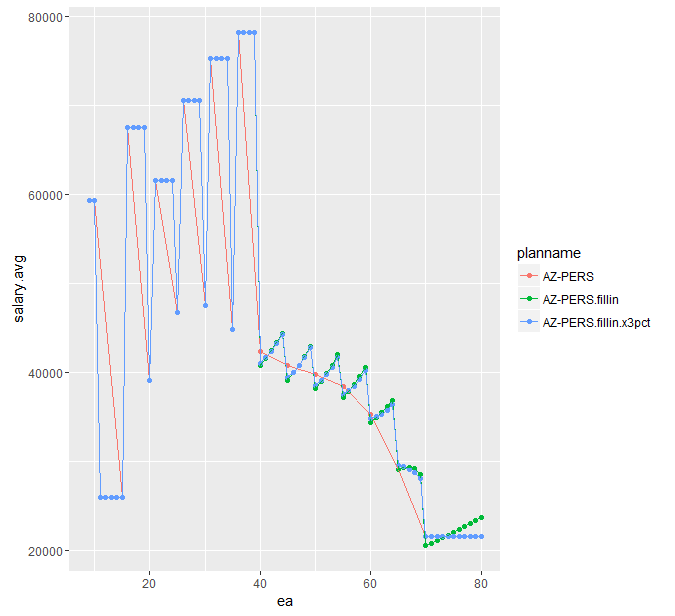
#### Average salary by age

Figure Avg salary by age: grouped; full scale; scale less 3% inflation

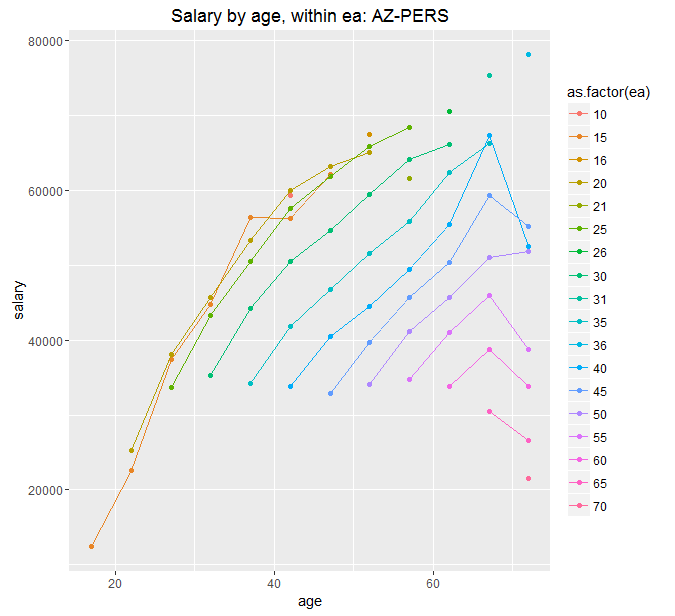


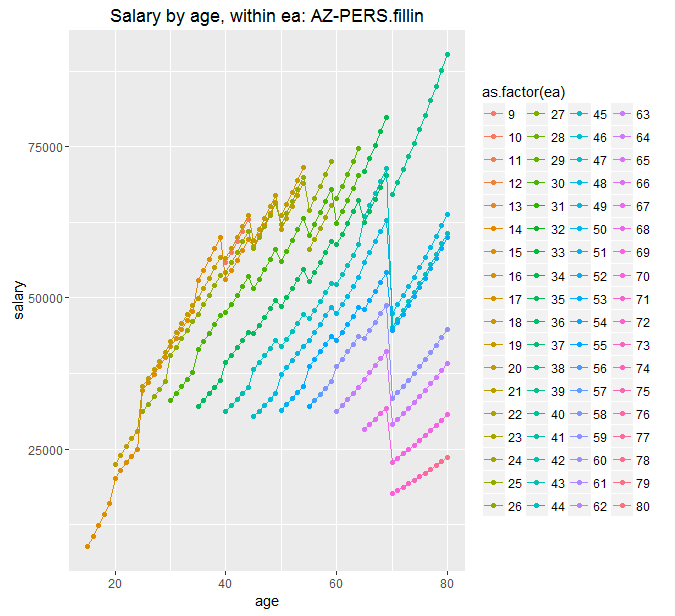
#### Average salary by ea

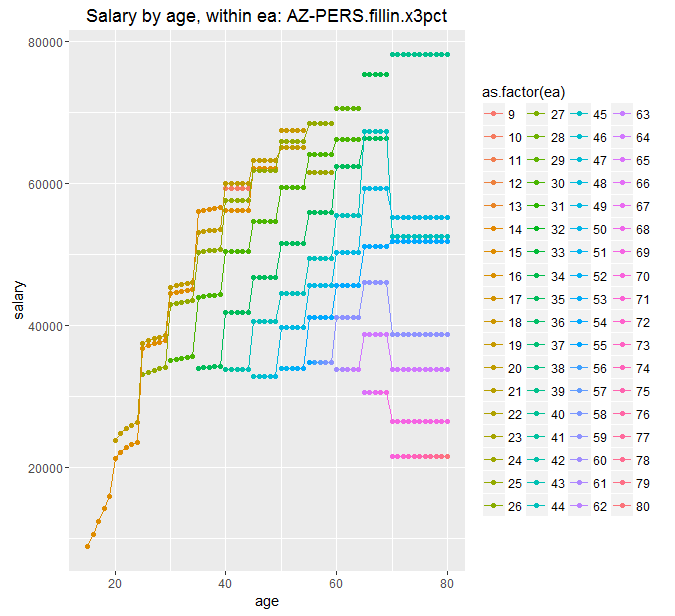
Figure 2 Avg salary by ea: grouped; full scale; scale less 3% inflation



#### Average salary by age, within ea







## Possible improvements

The most obvious improvement is to include the group lower and upper bounds with the data so that I don’t have to estimate them. I don’t think it will make much difference but I will get to it in the next few days.

After that, I think there are three other potential issues:

1. It might be possible to spread nactives in some other way than uniformly across allowable cells in a group. I am not sure what would be better. Maybe we could use a spline but I am worried that it could produce some odd results.
2. I think using the salary scale makes sense – I only use it within a specific age-ea group. It ensures that older workers within a group have higher salaries. But there certainly are cases in the data were salaries do not increase across age groups. We could again consider a spline.
3. We could bring in additional data as constraints – e.g, nactives and salary for individual ages, if avail in the ea. This would be a lot more work but might be worth it for the 5 plans we do in detail.

If you have ideas for improvement please let me know.