A Unique Identifier for Commercial Crewmember License Data

CFEC Report No. 08-1N
January 2008

Prepared by Cathy Tide

Alaska Commercial Fisheries Entry Commission
8800 Glacier Highway, Suite 109
P.O. Box 110302
Juneau, Alaska 99811-0302
(907) 789-6160

OEO/ADA Compliance Statement

The Commercial Fisheries Entry Commission is administratively attached to the Alaska Department of Fish and Game (ADF&G).

ADF&G administers all programs and activities free from discrimination based on race, color, national origin, age, sex, religion, marital status, pregnancy, parenthood, or disability. The department administers all programs and activities in compliance with Title VI of the Civil Rights Act of 1964, Section 504 of the Rehabilitation Act of 1973, Title II of the Americans with Disabilities Act of 1990, the Age Discrimination Act of 1975, and Title IX of the Education Amendments of 1972.
If you believe you have been discriminated against in any program, activity, or facility please write:

- ADF&G ADA Coordinator, P.O. Box 115526, Juneau, AK 99811-5526
- U.S. Fish and Wildlife Service, 4401 N. Fairfax Drive, MS 2042, Arlington, VA 22203
- Office of Equal Opportunity, U.S. Department of the Interior, 1849 C Street NW MS 5230, Washington DC 20240.

The department's ADA Coordinator can be reached via phone at the following numbers:

(VOICE) 907-465-6077
(Statewide Telecommunication Device for the Deaf) 1-800-478-3648
(Juneau TDD) 907-465-3646
(FAX) 907-465-6078

For information on alternative formats and questions on this publication, please contact the following:

Commercial Fisheries Entry Commission (CFEC)
Research Section
8800 Glacier Highway, Suite 109
P.O. Box 110302
Juneau, Alaska 99811-0302
(907) 789-6160 *phone*
(907) 789-6170 *fax*

Abstract

The licensing system utilized by Alaska Department of Fish and Game for issuing commercial crewmember licenses is practical; it enables participants in the commercial fisheries of Alaska to obtain licenses easily and across the state, including in remote locations. The licenses are issued annually, with no information linking a license holder from one year to licenses held in other years. Despite this, the Commercial Fisheries Entry Commission is interested in tracking individual crew license holders through time and matching license holders to Commercial Fisheries Entry Commission permit information when possible. This report summarizes a process which attempts to identify license records in the commercial crewmember data that represents the same individual and which assigns a unique identification number to the individual. The advantages and disadvantages of the identification numbers developed through this process are examined. License data associated with each identification number are evaluated, which highlights some of the obstacles overcome in assigning identification numbers to commercial crewmember data.

1.0 Introduction

Interest in commercial crewmember data has been growing. There is interest in studying commercial fishing employment trends, tracking crewmember participation in particular fisheries, and in light of potential allocation issues, developing histories for crewmembers based on participation in fisheries. The primary source of information about commercial crewmembers arises from commercial crewmember licenses issued by the Alaska Department of Fish and Game (ADF&G). According to ADF&G regulations, a person is required to obtain a commercial crewmember license in order to participate in commercial fishing in any waters of Alaska, if they do not already hold a valid Commercial Fisheries Entry Commission (CFEC) interim-use or limited entry permit card (5 AAC 39.110). ADF&G has been issuing commercial crewmember licenses since 1988.[1] With the issuance of each license, certain data are collected about the license holder. These data are used in this study.

ADF&G Division of Administrative Services (DAS) and roughly 1,600 vendors across the state of Alaska sell hunting, trapping, sport fishing, and commercial crewmember licenses.[2] The sale of commercial crewmember licenses through vendors, rather than strictly ADF&G DAS, makes them readily accessible to persons interested in working in the fishing industry.[3] A crewmember can be licensed and able to work on short notice and in remote locations.

This licensing system was intended to make obtaining a crewmember license easy and quick for participants in the fishing industry. It was not designed to be a definitive source of demographic data on crew, or be a means for tracking crewmember participation. With the new interest in crewmember information, people want to use the crew data for this purpose, however. This raises the question whether the data can be successfully used in a way that was never intended.

CFEC, as well as other groups, may have similar questions about commercial crewmembers and crewmember license data. Even though not intended for this purpose, could the crew data be used for demographic analysis? Could license holders be followed through time? Would it be possible to identify individuals that have held crew licenses and also held interim-use or limited entry commercial fishing permits? CFEC obtained crewmember data from ADF&G for license years 1988 through 2006. The initial step towards answering these questions was to analyze the contents of ADF&G commercial crewmember data, and do a preliminary evaluation of its quality. The results of these initial analyses can be found in *Preliminary Examination of Commercial Crewmember License Data*.[4] The next step in answering these questions was determining whether unique individuals could be identified in the commercial crewmember data and if a unique identifier could be assigned. This report summarizes the process utilized in assigning unique identifiers to individuals in commercial crewmember data and contains a brief discussion on the merits and shortcomings of the identification process. Analysis of license data associated with each identification number illustrates the obstacles that needed to be overcome through the identification process.

The key points identified through this study include:

- A unique identification number was assigned to each individual based upon a combination of license records.
- The assignment of identification numbers were made despite the variability of responses found in the social security number, birth date, first name, and last name fields of the commercial crewmember data.
- As a result of the automated process used, and the complicated data involved, the assignment of identification numbers was not flawless; the process sometimes resulted in more than one individual with the same identification number, and sometimes a single individual was assigned more than one identification number.

---

[1] Commercial crewmember licenses were issued by the Alaska Department of Revenue between 1984 and 1987. According to an internal CFEC memo dated August 9, 2004, by Laura Joralemon, crewmember license data from 1984 through 1987 exist on floppy discs that may or may not be readable as data from the old WANG system that the Department of Revenue used. There are microfiche copies of the license data in report format. Paper copies of the licenses do not exist.

[2] Wright, Kristin. 2007. Personal communication. Alaska Department of Fish and Game; P.O. Box 115526, Juneau, AK 99811-5526.

[3] While there are roughly 1,600 vendors each year, vendor codes in the crew license data for non-voided licenses attribute only 236 to 352 vendors each year with selling commercial crewmember licenses between 1988 and 2006.

[4] Tide, Cathy. 2007. *Preliminary Examination of Commercial Crewmember License Data*. Commercial Fisheries Entry Commission, CFEC Report 07-7N, Juneau.

- Analyst judgment was required throughout the 'automated' process which might result in different groupings of license records being categorized as a single individual if this process was reproduced by a different analyst.
- Analysis of commercial crewmember data using the identification numbers generated in this process should be viewed cautiously. They are estimates, rather than definitive facts about crew.

2.0 Need for a Unique Identifier

Two of CFEC's main objectives in looking at commercial crewmember data are to track individuals through time, and to determine if crewmember license holders have also held CFEC interim-use or limited entry permits. Crewmember licenses are issued annually and as such there is no definitive information linking a license holder from one year to licenses held in preceding years. Each crewmember has a different license number each year. Crucial to either of CFEC's objectives is the identification of license records that represent the same unique individual. Unfortunately, commercial crewmember data do not contain a completely reliable unique identifier.

Social Security Numbers (SSN) which are typically thought of as a unique identifier, only appear on roughly two-thirds of the license records. In addition, within the 19 years of crewmember data analyzed herein, there are cases of different people using the same SSN. The SSN, birth date, and first and last name fields each contain information in the commercial crewmember data that could be used to help identify unique individuals. Unfortunately, initial review of the data revealed that none of the fields could be used independently as a reliable unique identifier for individuals. These four fields may be used in combination to help estimate when different records represent the same individual, however. Please refer to *Preliminary Examination of Commercial Crewmember License Data*, for a more in-depth description of each of these fields.[5]

The very nature of commercial crewmember data complicates any attempt to assign unique identification numbers. Licenses can be obtained from vendors across the state and are simply forms filled out by the applicant. There can be a lot of variability among responses that appear within the data for the same individual. For example, full names, nick names, abbreviations, and initials can be used; last names change (e.g., due to marriage or divorce); data can be omitted; and numbers or letters can be transposed. All of these differences can arise from information provided by the license holder. That variability may be compounded by the ADF&G DAS staff that is required to interpret handwritten license applications and data enter the information into the crew license database. The variability in responses, whatever the source, for the SSN, birth date, first name, and last name fields, lends further support for the necessity of an identification number that can bridge all of those differences.

This paper focuses on the attempt to use the SSN, birth date, first name, and last name fields in a somewhat automated process to identify records that represent the same individual. To illustrate this approach in assigning identification numbers, imaginary commercial crewmember data were created which are representative of the kind of data seen in the actual database, and exhibit some of the issues faced in estimating which crewmember license records represent the same individual. This same example will be utilized throughout this paper to illustrate several stages in the process. Table 1 contains fictitious commercial crewmember license data. Please note the 6 SSNs, 5 first names, 3 last names, and 7 birth dates.

[5] Tide, Cathy. 2007. *Preliminary Examination of Commercial Crewmember License Data*. Commercial Fisheries Entry Commission, CFEC Report 07-7N, Juneau.

Table 1. Example of Commercial Crewmember Data Encountered While Assigning Unique Identification Numbers.

Record	SSN[1]	First Name	Last Name	Birth Date
1	XX1-22-3333	Jennifer	Doe	05/12/1980
2	XX1-22-3333	Jennifer	Doe	05/12/1980
3	XX1-22-3333	Jen	Doe	05/12/1980
4	XX1-22-3333	Jennifer	Poe	05/12/1980
5	XX1-22-3333	Jen	Doe	05/12/1930
6	XX1-22-3333	Jennifer	Doe	01/01/1901
7	XX1-22-3333	Jennifer	Doe	
8	XX7-22-3333	Jennifer	Doe	05/12/1980
9		Jennifer	Poe	05/12/1980
10	XX1-33-2222	Jen	Doe	05/12/1930
11	XX1-22-3333	Brian	Doe	07/24/1974
12		Brian	Doe	07/24/1974
13	XX1-23-4567	John	Smith	10/04/1968
14	XX1-22-3338	Jenn	Doe	12/05/1980

[1] The first 2 numbers of the SSNs are masked with 'XX' so a real SSN is not used inadvertently in the example.

3.0 Assignment of a Unique Identifier

Before attempting to track individuals through time or conduct demographic analyses on commercial crewmember data, records that represent the same individual must be identified. Once records believed to represent the same person have been flagged, an identification number must be applied to each of these records. The following sections describe an attempt to identify and flag license records as the same person. Due to the complexity of the commercial crewmember data, many steps were required in the process.

3.1 Initial Identification Number Assignment

A unique identification number was assigned to each identical combination of SSN, birth date, first name, and last name if the combination appeared more than one time in the crew license data.[6] Each license record with that combination was assigned the same identification number. Since each record had the same information in the SSN, birth date, first name, and last name fields, confidence that they represented the same individual was high. Combinations of SSN, birth date, first name, and last name that appeared only one time in the crew license data were not assigned an identification number.

Table 2 provides an example of the initial identification number assignment. In the table, record 1 and record 2 were identified as the same individual because the contents of their SSN, birth date, first name, and last name fields were identical (shown in bold). As such, the identification number 10,001 was assigned to each. Each time a record is assigned an identification number in the example, they will be flagged with '*'. The other license data in the example were not assigned an identification number because the contents in their SSN, birth date, first name, and last name fields, *in combination*, did not exactly match any of the other records.

[6] 55,656 unique identification numbers were assigned in this step. To receive a unique identification number, the license records must have an SSN value that is 9 characters long. The characters must be numeric, and may not be filler values (i.e., 000-00-0000, 001-01-0001, or 999-99-9999). License records with blank SSN values could not be assigned a unique identification number in the initial step.

Table 2. Assignment of the Initial Identification (ID) Number Based on Repeated SSN, Birth Date, First Name, and Last Name.

Record	SSN[1]	First Name	Last Name	Birth Date	ID Number
1	**XX1-22-3333**	**Jennifer**	**Doe**	**05/12/1980**	10,001*
2	**XX1-22-3333**	**Jennifer**	**Doe**	**05/12/1980**	10,001*
3	XX1-22-3333	Jen	Doe	05/12/1980	
4	XX1-22-3333	Jennifer	Poe	05/12/1980	
5	XX1-22-3333	Jen	Doe	05/12/1930	
6	XX1-22-3333	Jennifer	Doe	01/01/1901	
7	XX1-22-3333	Jennifer	Doe		
8	XX7-22-3333	Jennifer	Doe	05/12/1980	
9		Jennifer	Poe	05/12/1980	
10	XX1-33-2222	Jen	Doe	05/12/1930	
11	XX1-22-3333	Brian	Doe	07/24/1974	
12		Brian	Doe	07/24/1974	
13	XX1-23-4567	John	Smith	10/04/1968	
14	XX1-22-3338	Jenn	Doe	12/05/1980	

[1] The first 2 numbers of the SSNs are masked with 'XX' so a real SSN is not used inadvertently in the example.
* indicates an ID number assigned during this step of the process.

3.2 Inclusion with Initial Identification Number

Of the license records not assigned an identification number during the process described above, it seemed likely that some actually represented individuals already identified, but for whom some records did not have identical values in all 4 of the fields used to assign the initial identification number. Subsequent steps of the process attempted to associate records without identification numbers to an apparent individual already identified and assigned an identification number on other records.

License records that matched those of an apparent individual based on identical values in select combinations of the SSN, birth date, first name, or last name fields were assigned that individual's identification number.[7] Matching license records in this fashion involved several combinations of data fields. The following three tables illustrate how licenses with a variety of data were aggregated with records that were assigned an identification number in the initial step. The records were aggregated first using the SSN and birth date combination, then the SSN, first name, last name combination, and then the birth date, first name, last name combination.

Table 3 demonstrates how license records from the example were identified as the same individual as Jennifer Doe, a license holder already assigned an ID number of 10,001. Records 3 and 4 match records 1 and 2 based on the SSN and birth date field combination (shown in bold). These 4 records appear to represent the same individual, despite variations in both the first name and last name fields, and as such, should have the same identification number. The identification number from records 1 and 2 was also assigned to records 3 and 4. Discrepancies, differences, or errors in either of the name fields were overlooked using this step.

[7] With 4 fields (SSN, birth date, first name, and last name) there are 11 possible combinations which take into account 2, 3, or 4 of the fields. The 11 combinations are: 1. **SSN, birth date, first name, last name**; 2. **SSN, last name, first name**; 3. SSN, birth date, last name; 4. SSN, birth date, first name; 5. **Birth date, last name, first name**; 6. **SSN, birth date**; 7. SSN, last name; 8. SSN, first name; 9. **Birth date, last name**; 10. Birth date, first name; 11. Last name, first name. With the exception of the SSN and last name and SSN and first name combinations, the remaining 9 combinations were examined at one or several stages of the process to see if they could be used to include additional license records without an ID in the identification of an individual, by matching them to other license records already assigned an ID number. Of those 9, only 5 were actually used in the process described in this report. Those 5 are bolded in the list above.

Table 3. Inclusion with the Initial ID Number Based on SSN and Birth Date.

Record	SSN[1]	First Name	Last Name	Birth Date	ID Number
1	**XX1-22-3333**	Jennifer	Doe	**05/12/1980**	10,001
2	**XX1-22-3333**	Jennifer	Doe	**05/12/1980**	10,001
3	**XX1-22-3333**	Jen	Doe	**05/12/1980**	10,001*
4	**XX1-22-3333**	Jennifer	Poe	**05/12/1980**	10,001*
5	XX1-22-3333	Jen	Doe	05/12/1930	
6	XX1-22-3333	Jennifer	Doe	01/01/1901	
7	XX1-22-3333	Jennifer	Doe		
8	XX7-22-3333	Jennifer	Doe	05/12/1980	
9		Jennifer	Poe	05/12/1980	
10	XX1-33-2222	Jen	Doe	05/12/1930	
11	XX1-22-3333	Brian	Doe	07/24/1974	
12		Brian	Doe	07/24/1974	
13	XX1-23-4567	John	Smith	10/04/1968	
14	XX1-22-3338	Jenn	Doe	12/05/1980	

[1] The first 2 numbers of the SSNs are masked with 'XX' so a real SSN is not used inadvertently in the example.
* indicates an ID number assigned during this step of the process.

With each step in this approach, the process becomes more inclusive. Now, in the example, instead of matching only Jennifer Doe in the name fields, later merges can be based on Jennifer Doe, Jen Doe, and Jennifer Poe. This helps overcome some of the variability seen in commercial crewmember data records for the same person.

Table 4 demonstrates how license records in the example are identified as the same individual as Jennifer Doe based on SSN, last name and first name. Record 5 has the same SSN, first name, and last name data as record 3 (shown in bold). As such, the identification number 10,001 is assigned to record 5. Records 6 and 7 match records 1 and 2 (shown in italics) and are therefore also assigned an identification number of 10,001. The discrepancy or error in the birth date field, the use of filler data (i.e., 01/01/1901), or absence of information, does not prevent the matching of license records to an individual already assigned an identification number, since information in the other fields matches. License records associated with the identification number 10,001 became even more inclusive following this merge. Not only are the names Jennifer Doe, Jen Doe, and Jennifer Poe associated with ID 10,001, but the birth dates 5/12/1980, 05/12/1930, and 01/01/1901 are also associated with it.

Table 4. Inclusion with the ID Number Based on SSN, First Name, and Last Name.

Record	SSN[1]	First Name	Last Name	Birth Date	ID Number
1	*XX1-22-3333*	*Jennifer*	*Doe*	05/12/1980	10,001
2	*XX1-22-3333*	*Jennifer*	*Doe*	05/12/1980	10,001
3	**XX1-22-3333**	**Jen**	**Doe**	05/12/1980	10,001
4	XX1-22-3333	Jennifer	Poe	05/12/1980	10,001
5	**XX1-22-3333**	**Jen**	**Doe**	05/12/1930	10,001*
6	*XX1-22-3333*	*Jennifer*	*Doe*	01/01/1901	10,001*
7	*XX1-22-3333*	*Jennifer*	*Doe*		10,001*
8	XX7-22-3333	Jennifer	Doe	05/12/1980	
9		Jennifer	Poe	05/12/1980	
10	XX1-33-2222	Jen	Doe	05/12/1930	
11	XX1-22-3333	Brian	Doe	07/24/1974	
12		Brian	Doe	07/24/1974	
13	XX1-23-4567	John	Smith	10/04/1968	
14	XX1-22-3338	Jenn	Doe	12/05/1980	

[1] The first 2 numbers of the SSNs are masked with 'XX' so a real SSN is not used inadvertently in the example.
* indicates an ID number assigned during this step of the process.

Table 5 shows how license records in the example are identified as the same individual as Jennifer Doe based on birth date, last name, and first name. Since record 8 has the same birth date, first name, and last name as records 1 and 2, record 8 was assigned an identification number of 10,001 (shown in bold). Record 9 matched record 4 (shown in italics) and record 10 matched record 5 (shown in gray) using this same field combination. The discrepancy or error in the SSN field, or the absence of data, does not prevent a match to license records already assigned an identification number since values match based on the 3 other fields. Please note that matches are now based on the names Jennifer Doe, Jen Doe, and Jennifer Poe and on the birth dates 05/12/1980, 05/12/1930, 01/01/1901, and blank. Also, there is now more than one SSN associated with the 10,001 identification number. The SSNs XX1-22-3333, XX7-22-3333, XX1-33-2222, and blank are all associated with an ID number of 10,001.

Table 5. Inclusion with the ID Number Based on Birth Date, First Name, and Last Name.

Record	SSN[1]	First Name	Last Name	Birth Date	ID Number
1	XX1-22-3333	**Jennifer**	**Doe**	**05/12/1980**	10,001
2	XX1-22-3333	**Jennifer**	**Doe**	**05/12/1980**	10,001
3	XX1-22-3333	Jen	Doe	05/12/1980	10,001
4	XX1-22-3333	*Jennifer*	*Poe*	*05/12/1980*	10,001
5	XX1-22-3333	Jen	Doe	05/12/1930	10,001
6	XX1-22-3333	Jennifer	Doe	01/01/1901	10,001
7	XX1-22-3333	Jennifer	Doe		10,001
8	XX7-22-3333	**Jennifer**	**Doe**	**05/12/1980**	10,001*
9		*Jennifer*	*Poe*	*05/12/1980*	10,001*
10	XX1-33-2222	Jen	Doe	05/12/1930	10,001*
11	XX1-22-3333	Brian	Doe	07/24/1974	
12		Brian	Doe	07/24/1974	
13	XX1-23-4567	John	Smith	10/04/1968	
14	XX1-22-3338	Jenn	Doe	12/05/1980	

[1] The first 2 numbers of the SSNs are masked with 'XX' so a real SSN is not used inadvertently in the example.
* indicates an ID number assigned during this step of the process.

3.3 Confirmation and Clean-Up

When license holders were initially declared the same person, based on all four criteria (SSN, birth date, first name, and last name) confidence was fairly high that they were truly the same individual. But in later steps, when the criteria was reduced to two or three of those fields, the chances of mistakenly calling different individuals the same person increased, and confidence that the people with the same identifier were truly the same person decreased. Using an automated process on data as complicated as this, it was inevitable that some inappropriate identification number assignments would occur. In hopes of reducing the frequency of this, a subset of the records assigned an identification number by each combination of fields was reviewed. If any one combination appeared to result in an unsatisfactory number of incorrect identification number assignments, that combination of fields was not used. While this resulted in a reduction in identification number assignments, this was preferable to increasing the number of inappropriate assignments.

Unfortunately, a visual inspection of records, after assignments were made, revealed that this process also appeared to result in some unique individuals being assigned more than one identification number. In the initial step, where repeated SSN, birth date, last name, and first name fields were assigned identification numbers, some persons had more than one SSN, birth date, last name, and first name combinations that were repeated in the license data. As a result, they were assigned more than one identification number. This became apparent when attempting to include additional records in the ID assignment based on two or three of the fields. A clean-up was performed each time it was revealed that license holders who appeared to be the same person had been assigned more than one identification number. This was done so the multiple assignments were not perpetuated through each subsequent step.

3.4 Second ID Assignment

After these initial rounds, 315,071 of the 512,432 (61.5%) valid crewmember licenses issued between 1988 and 2006, had been assigned an identification number that was ultimately based on the initial assignment to SSN, birth date, first name, and last name combinations that were repeated in the data. Since so many licenses remained without an identification number, and visual inspection indicated that there appeared to be many individuals repeated in the remaining data, a second set of identification numbers were assigned to birth date, first name, and last name combinations that were repeated in the data.[8] These 3-field combinations provide less confidence in differentiating unique individuals than the 4-field combinations; therefore identification numbers in a different range of numbers were assigned, so they could be differentiated from individuals identified based on all four fields.

Table 6 indicates that record 11 and record 12, of the example, were identified as the same individual because the contents of their birth date, first name, and last name fields were identical (shown in bold). As such, the identification number 20,001 was assigned to each. Note that this number falls in a different range than those assigned during the first assignment of identification numbers. It is also interesting to note that one of the Brian Doe licenses is associated with an SSN also associated with Jennifer Doe licenses, yet they appear to be different individuals. Examples such as this illustrate another reason why the SSN cannot be relied upon as the sole identifier for individuals in commercial crewmember data.

Table 6. Assignment of Identification Numbers Based on Repeated Birth Date, First Name, and Last Name.

Record	SSN[1]	First Name	Last Name	Birth Date	ID Number
1	XX1-22-3333	Jennifer	Doe	05/12/1980	10,001
2	XX1-22-3333	Jennifer	Doe	05/12/1980	10,001
3	XX1-22-3333	Jen	Doe	05/12/1980	10,001
4	XX1-22-3333	Jennifer	Poe	05/12/1980	10,001
5	XX1-22-3333	Jen	Doe	05/12/1930	10,001
6	XX1-22-3333	Jennifer	Doe	01/01/1901	10,001
7	XX1-22-3333	Jennifer	Doe		10,001
8	XX7-22-3333	Jennifer	Doe	05/12/1980	10,001
9		Jennifer	Poe	05/12/1980	10,001
10	XX1-33-2222	Jen	Doe	05/12/1930	10,001
11	XX1-22-3333	**Brian**	**Doe**	**07/24/1974**	20,001*
12		**Brian**	**Doe**	**07/24/1974**	20,001*
13	XX1-23-4567	John	Smith	10/04/1968	
14	XX1-22-3338	Jenn	Doe	12/05/1980	

[1] The first 2 numbers of the SSNs are masked with 'XX' so a real SSN is not used inadvertently in the example.
* indicates an ID number assigned during this step of the process.

3.5 Inclusion with Second ID

Of the license records not assigned an identification number during the process described so far, it seemed likely that some represented individuals identified in this latest step, but for whom some records did not have identical values in the birth date, first name, and last name fields. Attempts were made to match records without identification numbers to licenses which had been assigned an identification number based on repeated birth date, first name, and last name information. Records that matched based on selected combinations of two or three of the SSN, birth date, first name, and last name fields were assigned the identification numbers, from the second range of numbers, that were associated with the license data. The confirmation and clean-up process was handled in the same manner as was done with the initial identification numbers. Field combinations that appeared to introduce too many incorrect matches were not used, and individuals with multiple ID assignments

[8] 21,011 unique identification numbers were assigned in this step.

were corrected to just one identification number. After assignment of the initial and secondary identification numbers, and inclusion of matching records, 374,986 of the 512,432 (73.2%) valid crewmember licenses issued between 1988 and 2006 had been assigned an identification number that was ultimately based on the initial assignment to SSN, birth date, first name, and last name combinations that were repeated in the data or to birth date, first name, and last name combinations repeated in the data.

3.6 Third ID Assignment

Visual inspection of the remaining data indicated that there still appeared to be repeated individuals in the data not yet associated with an identification number. A third set of identification numbers were assigned to SSN, first name, and last name combinations that were repeated in the license data.[9] A third range of identification numbers was used so the basis of identification could be distinguished from earlier ones. Following this third assignment, records remained that had not yet been assigned an identification number. Some appeared to represent individuals identified in this third step but that did not contain identical values in the SSN, first name, and last name fields, however. The license records that matched license records with an identification number based on selected combinations of two or three of the SSN, birth date, first name, and last name fields were also assigned the individual's identification number. The confirmation and clean-up process was handled in the same manner as was performed on earlier ID assignments. The frequency of inappropriate identification number assignments was determined and reduced when possible and multiple ID assignments to unique individuals were corrected. Following the first three rounds of identification number assignment, 378,861 of the 512,432 (73.9%) valid crewmember licenses issued between 1988 and 2006 had been assigned an identification number.

3.7 Fourth ID Assignment

A fourth range of numbers was used in assigning identification numbers to SSN and birth date combinations that were repeated in the remaining commercial crewmember data.[10] Because these identifications were based on just two fields, confidence that these are truly the same individuals was somewhat lower than those based on three or four identical fields. Visual inspection of a subset of the assignments indicated that the same individual really appeared to be repeated in the data and correctly assigned a unique identification number, however. Following this fourth assignment of identification numbers, license records still remained with no identification number. Attempts to identify license records among these which might also represent the individuals identified in this fourth step were unsuccessful. The license records without identification numbers could not be matched to those assigned an ID based on the values in other combinations of fields.

3.8 Final ID Assignment

The remaining 123,867 license records (24.2% of 512,432 valid crewmember licenses) did not match to other records in the commercial crewmember data and as such, it was assumed that these remaining records represent unique individuals that only appear once in crewmember data between 1988 and 2006. Final identification numbers, in a new range of numbers, were assigned to each of these records.[11]

Table 7 shows that records 13 and 14, from the example, were each assigned an identification number in this final step since they were unable to match to any of the other records in the commercial crewmember data. The assigned identification numbers fall in this final range of numbers indicating that the license holders appear to be individuals who only appear once in the crew license data. In this example, John Smith appears to be one such individual. He does not appear to be the license holder found on any of the other license records. On the other hand, it appears that Jenn Doe, of record 14, might be the same person as Jennifer Doe, found in other records. There are slight differences in each one of the fields used to identify individuals, and as a result record 14 was assigned its own identification number through this automated process.

[9] 1,737 unique identification numbers were assigned in this step.
[10] 4,533 unique identification numbers were assigned in this step.
[11] 123,867 unique identification numbers were assigned during this step.

Table 7. Assignment of the Final ID Number to Individuals That Appear Once in Commercial Crewmember Data.

Record	SSN[1]	First Name	Last Name	Birth Date	ID Number
1	XX1-22-3333	Jennifer	Doe	05/12/1980	10,001
2	XX1-22-3333	Jennifer	Doe	05/12/1980	10,001
3	XX1-22-3333	Jen	Doe	05/12/1980	10,001
4	XX1-22-3333	Jennifer	Poe	05/12/1980	10,001
5	XX1-22-3333	Jen	Doe	05/12/1930	10,001
6	XX1-22-3333	Jennifer	Doe	01/01/1901	10,001
7	XX1-22-3333	Jennifer	Doe		10,001
8	XX7-22-3333	Jennifer	Doe	05/12/1980	10,001
9		Jennifer	Poe	05/12/1980	10,001
10	XX1-33-2222	Jen	Doe	05/12/1930	10,001
11	XX1-22-3333	Brian	Doe	07/24/1974	20,001
12		Brian	Doe	07/24/1974	20,001
13	**XX1-23-4567**	**John**	**Smith**	**10/04/1968**	50,001*
14	*XX1-22-3338*	*Jenn*	*Doe*	*12/05/1980*	50,002*

[1] The first 2 numbers of the SSNs are masked with 'XX' so a real SSN is not used inadvertently in the example.
* indicates an ID number assigned during this step of the process.

4.0 Shortcomings of the Unique Identification Number

There is no system to link individual license holders from year to year. Critical data fields that could be used to identify an individual are not consistent across time. This report describes an attempt to identify individuals in the crew license data and assign an identification number to each of their licenses, after the fact. This section describes shortcomings in the method used for identifying license records that represent the same individual, and for assigning a unique identification number to all of the crew license records for that person.

Despite attempts to assign identification numbers to unique individuals, this process sometimes resulted in different people receiving the same identification number. In an attempt to minimize this, identification number assignments were reviewed throughout the process. In some cases, the number of assignments was large and it was not practical to review each assignment. In these cases a subset of the data was reviewed. Despite these efforts, a visual inspection of the resulting assignments suggests that there are still cases where more than one person received the same identification number.

Throughout this process attempts were made to match records without identification numbers to those records with identification numbers if they shared information in selected combinations of 2 or 3 of the SSN, birth date, first name, or last name fields. Visual inspection revealed that matching by some field combinations (like the birth date and first name field combination) resulted in the inclusion of license records that did not appear to be the same person. The choice was made to err on the side of caution and not use these combinations. As a result, some erroneous matches were not made, but any appropriate matches gained from the field combination were also lost. Despite taking this precaution, records which do not appear to represent the same person may have been assigned the same identification number.

Another unintended but possibly unavoidable result of this process was that sometimes an individual was inadvertently assigned more than 1 identification number. In many cases this is due to the nature of the data. There was considerable variability in the data provided by license applicants, ADF&G DAS staff was required to analyze handwriting, and error could be introduced through data entry. Records for the same individual may not be associated with each other, because the information available was different enough as to make the association impossible through an automated process.

While the prior problems with the identification numbers are disappointing, they are not surprising considering the nature of the data being evaluated and the attempt to use an automated process to help group records and

assign identification numbers. What is most disheartening about the identification number assignment is the belief that it is not reproducible. Despite trying to automate this process, analyst judgments were required at each step of the process. If repeated, the results might vary with each analyst attempting to reproduce the results. Each could potentially end with some different groupings of license records being categorized as a single individual.

And finally, since there are shortcomings with these identification assignments, all analysis and conclusions drawn from these data should be viewed with caution. Slightly different values, trends, and results may result from the exact same commercial crewmember data if a different approach is taken for assigning an identification number. Any results generated from commercial crewmember data, based on these unique identification number assignments, reflect the approach and the judgments the analyst had to make throughout the process. Therefore results generated from these data should be viewed cautiously as rough estimates rather than definitive answers.

5.0 Advantages from the Unique Identification Number

In spite of the shortcomings discussed in the preceding section, there have been positive benefits gained by assigning these identification numbers to commercial crewmember license data. Primarily, the effort has identified individuals who appear in the data multiple times but whose license records otherwise may not have been easily associated with each other due to variability or lack of information found in certain data fields. In addition, the process appears to have distinguished license holders which have the same response in a data field – such as SSN – but whose responses in other fields suggest they are actually different persons.

People can be tracked through time using the identification number. SSNs which are typically thought to act as a unique identifier for individuals were not collected in 1995, 1996, and 1997. With a unique identifier assigned to license records, one can evaluate participation across this time period. A relatively low percentage of license records contain SSN information in 2001, and the unique identifier allows one to track participation across that year as well. Without the unique identifier, roughly 4 years of data are lost in which individual participants can be tracked. The missing 4 years, out of a 19 year time series, represent a considerable portion of time (21.1%).

The ability to track individuals across time makes it possible to conduct analyses on the number of years that people obtain commercial crewmember licenses. License longevity can be used to help distinguish professional crew from occasional participants. The unique identification numbers also allow more accurate demographic analysis. Analysis can reflect the number of people with licenses rather than the number of licenses, and can avoid the problems associated with double-counting individuals that obtain duplicate licenses or multiple 7-day licenses in a year.

The unique identifier can group data for data clean-up. All records for a particular identification number could be examined and corrected responses chosen for fields in which it would be appropriate. This also might require a considerable amount of analyst judgment but would result in uniform data for the same person across records. CFEC hopes to match crew license holders to individuals that have held CFEC interim-use or limited entry permits. This matching will most likely be performed using these same fields. The ability to correct number transpositions or misspellings could make the project more manageable.

Developing the process used to identify unique individuals and assign an identification number revealed other weaknesses in the historic data. For example, 65.8% of the duplicate licenses do not indicate the class code of the original license and a variety of filler data is used for birth date and SSN fields. While improvements have definitely been made in data collection in recent years, further improvements could be made to make future crew data more useful. The assignment of a unique identification number, in particular, which would be associated with each license issued to a person in a single year, and across years, would be a considerable improvement to crew data. That would allow a person to be tracked across time and would not require a retroactive effort to try to estimate when records represent the same person.

6.0 Analysis of the Unique Identification Number

A total of 202,795 unique identification numbers were assigned to 512,432 valid commercial crewmember licenses between 1988 and 2006.[12] Table 8 indicates the number of unique identification numbers found per license year in the commercial crewmember data. This represents an estimate of the number of unique individuals that obtained a commercial crewmember license in each year, from1988 through 2006. The number of valid licenses exceeds the number of unique identification numbers since an individual may hold more than one license in a license year.

Table 8. The Number of Unique Identification Numbers (IDs) per License Year, 1988-2006.

License Year	Valid Licenses[1]	Unique IDs[2]
1988	31,733	30,734
1989	32,750	31,466
1990	36,583	35,175
1991	36,738	35,245
1992	36,032	34,673
1993	33,056	31,726
1994	32,613	31,301
1995	30,788	29,778
1996	28,564	27,662
1997	27,306	26,530
1998	25,169	24,127
1999	25,075	24,022
2000	24,229	23,305
2001	20,475	19,781
2002	17,349	16,759
2003	18,166	17,469
2004	18,661	17,955
2005	18,719	18,042
2006	18,426	17,377

[1] CFEC obtained 528,802 license records for license years 1988 through 2006. 16,370 of those records were voided and 512,432 were considered 'valid.'
[2] The number of valid licenses exceeds the number of unique identification numbers since an individual may hold more than one license in a license year.

Several analyses were performed on the crewmember data once the identification numbers were assigned. These analyses provide further insight into the variability of responses found in certain fields in the commercial crewmember data and highlight the obstacles overcome in order to identify unique individuals.

Table 9 indicates the number of times each unique identification number was found in the commercial crewmember data between 1988 and 2006. Over half (61.1%) of the identification numbers only appear once in the data, while others appear up to as many as 41 times in the 19 years of crew license history. With the

[12] The commercial crewmember data that CFEC obtained for license years 1988 through 2006 contained 528,802 records. Of those records, 16,370 were voided. The remaining 512,432 license records are considered 'valid' records and used for this study.

Table 9. Number of Unique ID Appearances in Commercial Crewmember Data, 1988 – 2006.

Number of Appearances	Number of Unique IDs	Percent of Unique IDs
1	123,867	61.1%
2	25,027	12.3%
3	14,074	6.9%
4	9,467	4.7%
5	6,727	3.3%
6	5,032	2.5%
7	3,878	1.9%
8	3,065	1.5%
9	2,505	1.2%
10	2,005	1.0%
11	1,572	0.8%
12	1,253	0.6%
13	1,009	0.5%
14	827	0.4%
15	659	0.3%
16	511	0.3%
17	439	0.2%
18	315	0.2%
19	194	0.1%
20	101	0.1%
21	63	< 0.1%
22	55	< 0.1%
23	37	< 0.1%
24	27	< 0.1%
25	28	< 0.1%
26	9	< 0.1%
27	13	< 0.1%
28	13	< 0.1%
29	6	< 0.1%
30	2	< 0.1%
31	5	< 0.1%
32	3	< 0.1%
33	3	< 0.1%
34	2	< 0.1%
36	1	< 0.1%
41	1	< 0.1%

possibility of more than one person being assigned the same identification number, those with 25 or more appearances in the data were reviewed. In every case but one, all of the observations appeared to be the same individual (85 out of 86 identification numbers).[13] In general, the high number of repeated identification numbers in the data is likely due to individuals having licenses in several years, or holding multiple 7-day licenses, or being issued duplicate licenses in one or many years.

Examination of identification numbers that appear many times in the commercial crewmember data highlighted the fact that many individuals have multiple values for first name, last name, birth date, and SSN across the license record. The following tables look at the number of responses associated with those fields. Each table illustrates inconsistencies in crewmember data and the obstacles that had to be overcome to identify individuals in the data.

Table 10 looks at the number of different SSN responses that are associated with each identification number which appears in the crew license data more than one time in the period between 1988 and 2006. A majority of the unique identification numbers had more than one value for their SSN (61.3%) and three persons had as many as 8 different SSN values among their license records. A review of crew license data for identification numbers with a high number of SSN values confirmed that all but one appear to reflect a single individual (37 of 38 identification numbers).[14] The high number of different values in the SSN field can result from blanks, filler data, and crewmember responses. The responses may contain mistakes submitted by the license applicant, number transposition introduced during data entry, or handwriting misinterpretation. Because individual people can have so many different SSN responses among their license records, the SSN field alone cannot act as a unique identifier for individuals.

[13] One identification number appeared to consist of 2 separate individuals. In that case, 26 of the records appeared to reflect one individual and 1 record a second individual.

[14] One identification number appeared to consist of 2 separate individuals. In that case, the 2 people used the same last name, identical birth dates, and also used identical addresses. The first names and middle initials used do not indicate a simple reversal of first name and middle initial for one person, however.

Table 10. Number of SSN Responses Associated With Unique Identification Numbers, 1988-2006.[1]

Number of SSN Responses[2]	Number of Unique IDs	Percent of Unique IDs
1	30,556	38.7%
2	40,478	51.3%
3	6,716	8.5%
4	985	1.3%
5	155	0.2%
6	33	< 0.1%
7	2	< 0.1%
8	3	< 0.1%

[1] Only the 78,928 identification numbers found more than one time are included in this analysis. The 123,867 identification numbers which appear only once in the data are excluded.
[2] Blanks or filler data in the SSN field are each counted as an SSN response.

Table 11 also supports the contention that the SSN field does not serve as an adequate unique identifier for individuals in the commercial crewmember data. The table indicates the number of identification numbers associated with each unique SSN found in the 19 years of commercial crewmember data. There were 152,998 SSN responses in the commercial crewmember data. Of those, 97.1% were associated with only one identification number. The remaining 4,489 SSNs were linked to between 2 and 102,184 different identification numbers. The SSN associated with 102,184 different identification numbers was a blank response. The SSNs with between 6 and 312 unique identification numbers were each invalid SSN values with fewer than 9 digits, or strings of 0's in the area, group, or serial number. For example, the SSN response of '000-00-0000' was associated with 226 unique identification numbers and '000- - ' was associated with 30. A review of SSNs associated with 4 or 5 identification numbers revealed cases of truly different people using the same SSN. There are also cases where the SSN appears to represent a single individual, but the first name, last name, and birth date fields show such disparity that the records could not be linked to one another through the automated process utilized in this study.

Table 11. Number of Identification Numbers Associated with Unique SSNs in Commercial Crewmember Data, 1988-2006.[1]

Number of Unique IDs	Number of SSNs	Percent of SSNs
1	148,509	97.1%
2	4,277	2.8%
3	189	0.1%
4	14	< 0.1%
5	1	< 0.1%
6	2	< 0.1%
7	1	< 0.1%
16	1	< 0.1%
30	1	< 0.1%
226	1	< 0.1%
312	1	< 0.1%
102,184[2]	1	< 0.1%

[1] All responses found in the SSN field are included in this table, including a blank response, values that are fewer than 9 digits, contain non-numeric characters, or those with '000' as the area number, '00' as the group number, or '0000' as the serial number.
[1] These unique IDs all reflect a blank SSN.

Disparity in the birth date, first name, and last name fields were examined using the assigned identification numbers. Table 12 indicates the number of birth date responses associated with identification numbers that were found in the crew license data more than one time between 1988 and 2006. A vast majority of the identification numbers were only associated with 1 birth date (85.6% of 78,928 identification numbers). The remaining identification numbers had between 2 and 6 different values in the birth date field. Closer examination of the identification numbers with 5 or 6 birth date responses confirmed that the identification number did, in fact, appear to represent a single individual. The variety of birth dates stem from blank responses, filler values, and license applicant responses. The responses may contain mistakes submitted by the license applicant, like the reversal of the month and date information. Frequently seen is the correct month and date of birth, but the accidental use of the license year instead of the birth year. Number transposition introduced during data entry and handwriting interpretation may be responsible for additional variation among birth date information for an individual.

Table 12. Number of Birth Date Responses Associated With Unique Identification Numbers, 1988-2006.[1]

Number of Birth Date Responses[2]	Number of Unique IDs	Percent of Unique IDs
1	67,564	85.6%
2	10,167	12.9%
3	1,062	1.3%
4	123	0.2%
5	11	< 0.1%
6	1	< 0.1%

[1] Only the 78,928 identification numbers found more than one time are included in this analysis. The 123,867 identification numbers which appear only once in the data are excluded.
[2] Blanks or filler data in the birth date field are each counted as a birth date response.

In order to examine the variability of names linked to unique identification numbers, the combined information from the first name and last name fields were considered as one response. The majority (62.1%) of identification numbers which appeared more than one time between 1988 and 2006 had a single name combination. However, a considerable number had 2 or more name variations (37.9%). The name fields showed more variability than the SSN and birth date fields, with as many as 13 responses for one identification number. The use of full names, nick names, initials, abbreviations, name changes, and hyphenations for the same individual contributed to the variability. A circumstance seen quite often was the reversal of first name and middle initial. As with the other fields, this variability could be compounded through data entry error and handwriting misinterpretation. A visual review of the identification numbers with 8 or more name combinations revealed only one identification number that appeared to represent 2 individuals (1 of 27 identification numbers). This is the same identification number identified earlier which appeared 27 times in the crewmember data; 26 observations appeared to be one individual and one of the observations appeared to be a different person. Although there are instances, such as this, where more than one person was assigned a single identification number, the number of individuals that were uniquely identified despite all the possible name, birth date, and SSN values was encouraging.

Table 13. First and Last Name Combinations Associated With Unique Identification Numbers, 1988-2006.[1]

Number of First and Last Name Combinations[2]	Number of Unique IDs	Percent of Unique IDs
1	49,014	62.1%
2	21,394	27.1%
3	5,873	7.4%
4	1,771	2.2%
5	559	0.7%
6	205	0.5%
7	85	0.1%
8	14	< 0.1%
9	8	< 0.1%
10	3	< 0.1%
11	1	< 0.1%
13	1	< 0.1%

[1] Only the 78,928 identification numbers found more than one time are included in this analysis. The 123,867 identification numbers which appear only once in the data are excluded.
[2] Responses in the first and last name field were compressed to remove blank spaces and concatenated to each other to create the combinations examined here.

7.0 Summary

The licensing system utilized by ADF&G for issuing commercial crewmember licenses is practical in that it enables participants in the commercial fisheries of Alaska to obtain a license quite easily, even in remote areas of the state. The information collected with the sale of each commercial crewmember license has become the focus of growing interest. There is interest in studying commercial fishing employment trends, tracking crewmember participation in particular fisheries, and developing a system through which crewmembers' participation histories can be tracked.

CFEC is among the groups interested in commercial crewmember data. Our interest in the data focuses on tracking an individual crew license holder through time and answering such questions as: how many years do persons typically work as a crewmember? or, how many commercial crew license holders have also held CFEC interim-use or limited entry permits? The data collected from the sale of licenses were never intended for this purpose. Licenses are issued annually and as such, there is often no definitive information linking a license holder from one year to the next. Each crewmember has a different license number each year. CFEC hopes that the crew license data, although not intended for this purpose, can be used to track individuals through time. Estimating which observations in the data belong to the same person and assigning a unique identification number to those observations is essential for this to occur.

Because no single field in the existing commercial crewmember data serves adequately as a unique identifier, this analysis has attempted to identify unique persons using combinations of the SSN, birth date, first name, and last name fields. Once identified, the process assigned an identification number. The process was based on licenses with repeated information in all four fields and licenses with information that matched those based on combinations of two or three of the fields. Subsequent steps in the process identified individuals based on licenses with repeated information in the birth date, first name, and last name fields; the SSN, first name, and last name fields; and SSN and birth date fields. Additional licenses, that shared information based on two or three fields, were assigned the same identification number as the licenses with repeated information. At each step in the process, a subset of licenses identified for unique individuals were reviewed to confirm that they did in fact appear to be the same person.

The assumptions used in the computerized matching rules in this process led to some errors in identification number assignments. Despite attempts to assure identification numbers were assigned to unique individuals, there were cases where different people received the same identification number, or what appears to be a single person received more than one identification number. While unfortunate, these unfavorable outcomes were not all that surprising considering the assumptions needed to link large amounts of data in an automated process. Despite attempts to develop a strictly automated process for identifying unique individuals, analyst review and judgment were required throughout the process. Another analyst might make different judgment calls, resulting in some different identification number assignments. The reader should consider the identification number assignments as estimates and view reports from these assignments with suitable caution.

In spite of these difficulties, the process discussed herein appears to have succeeded in identifying individuals that appear multiple times and whose licenses could not otherwise have been easily associated with each other. In addition, the analysis found observations where two or more individuals shared the same information in some of the data fields. While this process and analysis did not provide perfect results, it provided a mechanism to link dissimilar records for the same individual so that a person's crew licensing could be tracked over time. Unique identification numbers were found up to as many as 41 times in the 19 years of crew data examined here. This process was able to link disparate observations despite the fact that some unique individuals had up to as many as 8 SSN responses, 6 birth dates, or 13 first and last name combinations.

Now that a unique identification number has been assigned to individuals in the commercial crewmember license data, more questions about commercial crewmember license holders can be addressed.